Novel computational methods using coded patient phenotypes to enhance disease gene identification

Saklatvala, Jake Robert

Awarding institution:
King's College London

Novel computational methods using coded patient phenotypes to enhance disease gene identification

Jake Saklatvala

1454769

Department of Medical and Molecular Genetics

September 2018

Thesis submitted to King's College London in fulfilment of the degree of Doctor of Philosophy

Declaration

I hereby declare that the work presented in this PhD thesis is my own.

Acknowledgements

I would firstly like to express my sincere gratitude to my supervisor Prof Michael Simpson, for his continued support and expertise throughout the past four years. It was a privilege to study under his supervision, and I am further grateful that his knowledge and dedication were rivalled by the patience and understanding he showed me throughout. I would also like to thank my second supervisor Prof Chris Mathew for his valuable advice and input into my project. I am also grateful to the organisations that funded this PhD - the Generation Trust and the Peter Stebbings Memorial Charity.

I would like to thank all members of the Simpson group for creating a friendly and constructive atmosphere where everyone is encouraged to share their work and help is always at hand. Particular thanks to Nick for taking the time to read my work and for his endless supply of useful advice and suggestions, as well as Christos and Ines for their support and thoughtful contributions. I would also like to thank Mark for his support in proofreading my thesis.

When I joined King's I was immediately welcomed by people from within the department and elsewhere, who I am glad to now call my friends. Thanks to Ines for her constant encouragement and support from the day I started, as well as being able to sit next to me for four years! Thanks to Christos for his words of wisdom, which were truly appreciated even if it seems like I rarely listened! Thanks to Carmen for her friendship and advice, especially during the final few months! Thanks to Duda for her support from the other side of the world! Thanks to the rest of the attendees of the Thomas Guy club on Fridays who made the end of the working week so enjoyable: Matt, Seth, Chronis, Laura, Diego, Dorita, John and

many more. I must also thank Sonny, who throughout my PhD has remained my closest friend.

I was also privileged to represent GKT (the finest football club in the world!) during my time at King's. I truly appreciate the bonds I formed with my teammates on and off the pitch, which helped me to unwind when work became particularly stressful.

I would finally like to thank my family for their unconditional love and encouragement throughout my life. I am extremely fortunate to have been raised by a mother and father who have made great sacrifices to help me get to where I am today. Thanks also to my sister, Hannah, my grandfather and my late grandmother, who have always been by my side.

Abstract

With the sequencing of the genomes of individuals with rare Mendelian disease becoming routine, there is an emerging challenge in identifying and quantifying similarity between individual's phenotypes to assist in the identification of commonalities in the genetic variation contributing to disease. Whilst it is relatively easy to assess genetic similarities between individuals, it is less trivial to assess phenotypic similarity due to the complexity of phenotypic information. One route to systematically estimate similarity between phenotypes utilises computational approaches applied to standardised machine-readable phenotypic descriptors, such as those in the Human Phenotype Ontology (HPO) or structured patient questionnaires. This thesis describes advances in the representation of clinical phenotypes in machine-readable controlled vocabulary within the context of genetic studies of both the diagnosis of monogenic disease patients, and common variant association analysis of severe acne subtypes. When using genome sequencing for the genetic diagnosis of individuals with rare Mendelian diseases, a virtual gene panel approach is often taken wherein only a curated list of genes suspected to cause a phenotype are considered. With the number of known monogenic disease-gene pairs exceeding 5,000, manual curation of personalised gene panels based on the entire human phenotypic spectrum is challenging. Methods have previously been developed that formalise the approach using the patient phenotype to generate candidate genes, requiring both patients and known disorders to be defined in standardised machine-readable terms. Work in this project has investigated the ways by which established phenotypic descriptions (OMIM free-text) can be further leveraged using simple quantification of disease

terms to gain a more nuanced description of known phenotypes with HPO terms, and how this helps to more efficiently generate candidate gene panels in real patient datasets. This project also examines the utility of extensive patient questionnaire records in patients with severe acne, enabling the identification of questionnaire response stratified subtypes of acne for use in downstream investigations seeking to identify new genetic determinants of the disease.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

ACMG	American College of Medical Genetics
BOQA	Bayesian Ontology Query Algorithm
CADD	Combined annotation dependent depletion
CNV	Copy number variation
CRF	Case report form
DAG	Directed acyclic graph
DBSCAN	Density-based spatial clustering of applications with noise
DDD	Deciphering Developmental Disorders
DDG2P	Developmental Disorders Genotype-Phenotype Database
DLQI	Dermatology Life Quality Index
eQTL	Expression quantitative trait loci
ESP	[NHLBI “Grand Opportunity”] Exome Sequencing Project
ExAC	Exome Aggregation Consortium
FORGE	Finding of Rare Disease Genes [Canada]
GO	Gene Ontology
GSTT	Guy’s and St. Thomas’
GWAS	Genome-wide association study
HGMD	Human Gene Mutation Database
HPO	Human Phenotype Ontology
IC	Information content
ICD	International Classification of Diseases
LOF	Loss-of-function

MAF	Minor allele frequency
MAP	Mean average precision
MICA	Most informative common ancestor
MPO	Mammalian Phenotype Ontology
NCBO	National Center for Biomedical Ontology
OMIM	Online Mendelian Inheritance in Man
OMIM PS	OMIM phenotypic series
PCA	Principal component analysis
PhenIX	Phenotypic Interpretation of eXomes
PheWAS	Phenome-wide association study
ROC	Receiver operating characteristic
SIFT	Sorting intolerant from tolerant
SNOMED CT	Systematized Nomenclature of Medicine -- Clinical Terms
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
t-SNE	T-distributed Stochastic Neighbour Embedding
UPD	Uniparental disomy
VSM	Vector space model
VUS	Variant of unknown significance
WES	Whole exome sequencing
WGS	Whole genome sequencing

Chapter 1 - General Introduction

1.1 Identification of disease-causing genetic variation

1.1.1 Genetic disease

Genetic diseases are caused by a change in an individual's DNA sequence which increases the likelihood of developing the disease. The extent to which genetic variation is causative of disease can range from that of fully penetrant monogenic disorders, where a single sequence variant is sufficient to cause the disease, to common complex diseases, where there are often multiple individual genetic variants that influence susceptibility to disease, often in addition to and in combination with lifestyle and environmental factors. To generalise, monogenic disease is usually rare, severe and frequently presents in childhood/adolescence, whereas common complex diseases often present in adulthood or later life. Complex diseases are generally more common than monogenic diseases because causative variants have lower effects on disease risk and are therefore under less selective pressure than variation that would cause severe monogenic disease, and can elevate in frequency within the population (Blekhman et al., 2008).

To make a confident statement about the role of a specific allele or group of alleles with similar implication in a disease phenotype requires either the recurrent observation of these alleles within a cohort of individuals with the phenotype, or the identification of a known relationship between the phenotype and allele from within all documented causal relationships that have been discovered. Relationships between disease phenotypes and alleles at genomic loci can be discovered though either linkage mapping (followed by candidate gene sequencing) or association studies.

1.1.2 Linkage mapping

Linkage mapping uses genome-wide markers to reveal chromosomal segments shared between individuals within pedigrees in which a trait also segregates, enabling statistical testing for non-random associations between marker loci and the disease trait. Linkage mapping is most suitable for identifying loci where highly penetrant genetic changes cause disease but do not impact on reproductive fitness, hence familial segregation of the disease is observed. Roughly a third of Mendelian disease genes have been discovered using linkage mapping and candidate gene resequencing (Bamshad et al., 2011; McKusick, 2007). However, linkage mapping is ineffective with locus heterogeneity, small pedigrees (which are likely in disorders with diminished reproductive fitness or caused by *de novo* mutation) and reduced penetrance (Kaiser, 2010).

1.1.3 Association studies

Association studies use genotype data from individuals to identify alleles that are present in greater or lower frequency in those with the phenotype compared to unaffected controls. Association studies do not require groups of individuals from within the same pedigree, but rather require adequate numbers of cases and controls for statistical power and an appropriate method of genotyping for the genetic study undertaken (i.e. for the identification of either highly penetrant variation in rare disease, or variation with smaller effect size in complex disease). A key element of association studies is that they require the identification of individuals with the same phenotype, as well as the identification of controls without the phenotype (which must be appropriately stratified). In rare disease it is particularly challenging to recruit individuals with similar phenotypes due to the scarcity of cases – it is therefore crucial that observed relationships between rare phenotypes

and allelic variants are recorded as a reference for individuals existing outside of these studies, which enables them to receive a molecular diagnosis without having to "rediscover" disease genes that have already been identified.

1.1.4 Association studies in rare monogenic disease

1.1.4.1 Unit of association

For rare monogenic disease, under the assumption of complete penetrance, putative causative alleles must be exclusively observed in individuals with the disease. However, variation at multiple positions within a gene may give rise to the aberrant gene function that causes the phenotype, so variant-based testing often lacks power. As a result, gene level testing is often employed, where genome-wide variation is filtered based on the expected properties of gene-disrupting variants – variation can then be collapsed to gene level to test for significant increases in putative pathogenic variation in cases compared to controls.

1.1.4.2 Whole genome and whole exome sequencing

Using linkage mapping, the power to identify causative variants in rare disorders is limited in instances where there is either a small number of cases, locus heterogeneity or a rare disorder arising spontaneously through *de novo* mutation (Ng, Buckingham, et al., 2010). Advances in whole genome sequencing technology have resulted in sequencing cost declining over four orders of magnitude compared to Sanger sequencing (Metzker, 2010; Wetterstrand, 2016), and as a result whole genome and whole exome sequencing have become widely used for the identification of disease genes in rare monogenic diseases, enabling genotypes at almost every locus in the genome/exome to be tested for association with disease status. Whole genome sequencing (WGS) entails the sequencing of a patient's entire genome at sufficient depth (At least 20X is required to accurately call a

heterozygous variant (Choi et al., 2009; Depristo et al., 2011)), but a more cost-effective option is available in whole exome sequencing (WES). WES selectively targets the protein coding regions, which constitute ~1.5% of the genome. It is estimated that ~85% of mutations with large effects on disease-related traits reside in the exome (Majewski, Schwartzentruber, Lalonde, Montpetit, & Jabado, 2011) and that exonic mutations cause the majority of monogenic diseases (Kuhlenbäumer, Hullmann, & Appenzeller, 2011). Due to the excess of disease-related variation contained within this small fraction of the genome, WES has become a powerful cost-effective alternative to WGS (Boycott, Dyment, Sawyer, Vanstone, & Beaulieu, 2014).

1.1.4.3 Filtering potential monogenic disease-causing variants

The exome sequencing of an individual identifies ~20,000 sequence variants on average (Hoischen et al., 2010; Musunuru et al., 2010; Ng et al., 2009), so the challenge of identifying a single disease-causing variant (or pair of variants in recessive disease) involves a substantial reduction of information. There are several filters that can be applied that help to achieve this based on theorised and observed properties of monogenic disease-causing variants:

- **Non-rare variants:** Large datasets of unaffected individuals are often used to estimate the population frequency of variants and remove variants with an allele frequency above a certain threshold. This includes datasets such as 1,000 genomes (n = 2,504) (1000 Genomes Project Consortium et al., 2015), the NHLBI Exome Sequencing Project (ESP) (n = 6,503) (Fu et al., 2013) and the Exome Aggregation Consortium (ExAC) (n = 60,706), later known as the Genome Aggregation Database (gnomAD) (n = 123,136 exomes + 15,496 genomes) (Lek et al., 2016). The larger the sampled

population and the more it reflects the ancestry of the patient, the more accurate the allele frequency estimate will be. The appropriate frequency threshold to use is dictated by the suspected inheritance model of the disease, prevalence of the disease in the population and the expected penetrance of the variant. Some variants annotated as disease-causing of autosomal dominant disorders in the Human Gene Mutation Database (HGMD) (Stenson et al., 2009) have been found to be present in healthy exomes, which potentially indicates reduced penetrance and therefore frequency filters may be relaxed accordingly. However, this may also be indicative of sequencing errors or false-positive entries in HGMD (Winand et al., 2014). It is also useful to filter by allele frequency observed in databases that comprise individuals sequenced at the same facility (using the same platforms(s)) – this firstly provides another estimate of allele frequency but more importantly can reveal common variants called due to sequencing/calling artefacts exclusive to the facility/platform used.

- **Zygosity:** Variants can be filtered on zygosity (homozygous/heterozygous) to reflect the suspected mode of inheritance (if known). If a recessive model is assumed, only variants that are homozygous or consistent with compound heterozygosity would be considered. To confirm that compound heterozygous mutations exist in *trans*, sequence data from the parents must be used to establish that the patient inherited one variant from each parent.
- **Functional consequence:** Even when sequencing is not limited to the exome, variation is often filtered to the coding region of the genome, in which functional interpretation is much better understood (Goldstein et al., 2013).

- **Synonymous variants:** Synonymous variants typically constitute ~50% of total called exonic single nucleotide variants (SNVs) (Bamshad et al., 2011), and these are commonly removed because they do not affect protein sequence. However, there is increasing evidence that synonymous mutations can affect protein expression, conformation and function, and it is estimated that 5-10% of genes contain at least one damaging synonymous mutation (Sauna & Kimchi-Sarfaty, 2011).
- **Pathogenicity prediction:** Missense mutations typically comprise just under 50% of called SNVs, and these can have vastly varied effects on protein function based on the nature of the amino acid change and its location within the protein sequence. This has encouraged the development of tools that predict whether an SNV is damaging (alters the normal levels or biochemical function of a gene or gene product) or deleterious (reduces the reproductive fitness of carriers and would thus be targeted by purifying natural selection) based on a model incorporating one or many annotations. Although the estimation of pathogenicity based on these predictions can help prioritise variant data and provide additional lines of evidence for whether a mutation is benign or deleterious, it is not recommended that they are considered alone (or using consensus of multiple scores) as a strong line of evidence due to the uncertain relationship between pathogenicity and annotation-derived damaging/deleteriousness scores (Kircher et al., 2014; D. G. MacArthur et al., 2014).

- **SIFT** (Sorting Intolerant from Tolerant) (Kumar, Henikoff, & Ng, 2009) assesses amino acid substitutions based on the premise that important positions in a protein sequence are evolutionarily conserved. SIFT conducts a protein sequence homology search to score all possible amino acids at a given position, and changes from highly conserved residues are scored as more deleterious.
- **PolyPhen2** (Adzhubei, Jordan, & Sunyaev, 2013) makes use of a classifier trained to predict whether genetic changes are damaging based on several annotations for known disease-causing mutant alleles (and their wild-type counterparts). Annotations are based on both sequence and protein structure.
- **CADD** (Combined annotation dependent deletion) (Kircher et al., 2014) makes use of an SVM classifier based on 63 annotations, trained to distinguish between 14.7 million high-frequency human-derived alleles (largely fixed in the human lineage) and an equivalent number of simulated *de novo* mutations. CADD scores are then computed for all 8.6 billion possible human SNVs. **DANN** (Quang, Chen, & Xie, 2015) utilised the same annotations and training set but uses a deep-learning framework, which outperforms CADD in their benchmarking.

In the absence of sequencing data from large family pedigrees to check whether a particular variant segregates with disease status, it is unlikely that these aforementioned variant filters will sufficiently narrow down towards the causative variant in a single individual, due to the amount of non-synonymous variation possessed (Bamshad et al., 2011) and rare LOF mutations harboured by each

individual (Daniel G. MacArthur et al., 2012). Depending on whether WES is undertaken to provide a genetic diagnosis or to discover novel genes, there are different strategies that enable identification of the single causative variant.

1.1.5 Monogenic diagnosis using WES

Once the causative relationship between gene and phenotype has been robustly defined by either linkage or association, it is possible to diagnose an individual based upon their phenotype and genotype. WES is now commonly used for the genetic diagnosis of individuals with rare disorders, particularly in instances where the number of possible candidate genes is sufficiently high that WES would be more financially viable than panel tests. This is the case when the patient's clinical phenotype is sufficiently unclear that one of multiple known disorders could be responsible, or when a particular disorder has a large number of potential causal genes (Bamshad et al., 2011). The use of exome sequencing has been found to be more effective than gene panel tests in achieving molecular diagnoses for patients (Neveling et al., 2013).

Often, standard filtering approaches are combined with virtual gene panels – prespecified lists of known causative genes that are prepared for specific phenotypic areas – and only variants in the panel are considered as potentially causal. Virtual panel size is variable – highly specific panels with few genes can be used, which may result in the exclusion of the disease gene, and some panels contain over 1,000 genes (such as the DDG2P developmental gene panel (Wright et al., 2015)) which may not be sufficiently specific to the patient phenotype and will result in more variants to be manually evaluated. Panels can be selectively augmented based on clinical evaluation of the patient phenotype but the number of monogenic disease-gene pairs has now surpassed 5,000 (Amberger, Bocchini,

Schiettecatte, Scott, & Hamosh, 2014), so it is becoming less feasible to construct personalised virtual panels for every patient.

Following the sharp decrease in the cost of sequencing (Wetterstrand, 2016), it has become feasible in a clinical diagnostic setting. As a result, a number of diagnostic sequencing projects have been undertaken, either focussed around one phenotypic area (de Ligt et al., 2012; Y Yang et al., 2014; Yaping Yang et al., 2013) or across multiple broad phenotypic areas (H. Lee et al., 2014; Sawyer et al., 2016; Taylor et al., 2015; Wright et al., 2015). These studies either utilised a primary gene panel to filter variants or involved manual review of all rare deleterious variants with respect to patient phenotype. Diagnostic yields were consistently reported to be between 25 and 30% (Schwarze, Buchanan, Taylor, & Wordsworth, 2018).

1.1.6 Monogenic gene discovery using WES

If the patient clinical evaluation doesn't match any known phenotypes, or diagnosis using virtual panels is unsuccessful, it is possible that the disease-causing variant resides in a gene previously undescribed to be causative of the phenotype. Novel gene discovery studies usually rely upon sequencing multiple unrelated patients with the same phenotype in order to identify the causal gene. After applying the filtering strategies such as removal of synonymous non-rare variants, the genes of the remaining variants are intersected. Studying unrelated patients minimises the number of identical-by-descent sequence variants. This method has been successful (Gilissen et al., 2010; Lalonde et al., 2010; Ng, Bigham, et al., 2010; Ng, Buckingham, et al., 2010; Simpson et al., 2011) although it requires a sufficient number of patients, and that the disorder has minimal genetic heterogeneity. Due to the rarity of monogenic disorders, it can be difficult to recruit patients to genetic

studies, as they may exist across different regions and may not be united by a single healthcare record system through which similarities can be identified.

If pedigree information is available, it can be utilised to narrow the search space for the causative variant. The causative variant is expected to segregate with the phenotype status within the family, so variants present in unaffected individuals can be discarded.

1.1.7 Validation of disease-causing variants

The description of a novel causal relationship between a gene and phenotype crucially requires statistical evidence that a variant is significantly enriched in cases compared to controls. Furthermore, by using sequence data from the family, statistical evidence can be provided that the variant is co-inherited with disease status. The variant must also be rare in a large population cohort with similar ancestry to the patient(s) (which will likely be true due to filtering). (D. G. MacArthur et al., 2014)

Informatic evidence in the form of pathogenicity prediction tools are considered useful additional lines of evidence based on sequence conservation and/or other various annotations of sequence context or resultant protein structure. This evidence should be supplemented by experimental evidence, which can be obtained by proving that the variant alters levels, splicing or normal function of the affected gene product. Either patient cells or an appropriate *in vitro* model system can be used. It is also useful to observe that the phenotype can be recapitulated by introducing the variant into a model system, as well as rescued by introducing the WT gene product. This requires a definition of a phenotype that is consistent with the human disease. (D. G. MacArthur et al., 2014)

For the genetic diagnosis of an individual, the specific statistical evidence required to associate a genetic variant with a clinical phenotype is less clear. If sequence data is not available for family members (both affected and unaffected), evidence cannot be gathered to assert that a particular variant segregates with disease status within the family. The requirement to make such genetic diagnoses requires the matching of a patient's phenotype to the phenotype of a known monogenic disorder (assuming it has been discovered) to a degree of confidence, as well as a rare deleterious variant in the relevant gene. To make such comparisons between individuals and known phenotypes, it is required that all discovered known phenotypes are recorded, as well as all known phenotype-gene causal relationships.

1.1.8 Documented phenotype-gene pairs

The documentation of known causal relationships between genes and phenotypes within databases is an increasingly useful clinical resource for the investigation of possible pathogenicity of candidate variants with respect to patient phenotypes.

OMIM (Online Mendelian Inheritance in Man) (Amberger et al., 2014) is a database which comprehensively documents human genes and genetic phenotypes. It is an online continuation of McKusick's *Mendelian Inheritance in Man* (MIM) (McKusick, 1966) and it contains information on over 15,000 genes and 6,000 Mendelian phenotypes (and an additional 2,000 phenotypes with suspected Mendelian basis). Over 5,000 monogenic phenotype-gene relationships are recorded, and clinical synopses of Mendelian phenotypes are included. Although precise medical terminologies are used, it is not a standardised form of phenotypic language. There is also a large amount of free-text containing phenotypic information.

Orphanet (Maiella, Rath, Angin, Mousson, & Kremp, 2013) is another database documenting rare disease (including infectious diseases). It contains various information on disease prevalence, inheritance, genes involved, as well as various health care resources and research activities on each disease. It also contains free-text descriptions of the phenotypic characteristics and categorisation of disease by features.

Variant databases **HGMD** (Stenson et al., 2009) and **ClinVar** (Landrum et al., 2018) list genetic variants reported to be involved in human disease, along with limited information on the phenotypes caused by each variant (or mappings to phenotypes within OMIM or Orphanet).

1.1.9 ACMG variant guidelines

The American College of Medical Genetics (ACMG) has published guidelines on the clinical interpretation of sequence variants identified in patients (Richards et al., 2015). These have been widely adopted, and the UK NHS has recommended implementation across genetic diagnostic laboratories testing for rare diseases and familial cancers (Ellard et al., 2017). The guidelines use a five-tier system (Table 1) to describe variants based on supporting evidence from a range of evidence sources, such as population minor allele frequency (MAF), computational predictions, functional study, segregation throughout family members.

Table 1: Five-tier variant classification from ACMG guidelines. Values in *% certainty pathogenic* are approximate recommended numerical meanings for each term rather than rigours calculations of probability.

Class	*Description*	*% certainty pathogenic**
5	Pathogenic	~100
4	Likely pathogenic	≥90
3	Unknown significance	10-90
2	Likely benign	≤10
1	Benign	~0

Two separate classifications for benign and pathogenic assertions have been designed (Figure 1), as well as rules for combining multiple sources of evidence into the final five-tier classification (Table 2). Although the term “likely” doesn’t confer a specific likelihood of a variant being either pathogenic or benign, guidelines recommend that “likely” corresponds to 90% certainty of either benign or pathogenic classification. The guidelines consider classifications of both pathogenic and likely pathogenic to be evidence that can be used in clinical decision making. Typically, variants classified as both pathogenic and likely pathogenic are considered to have sufficient evidence for health-care providers to use the molecular testing information in clinical decision making.

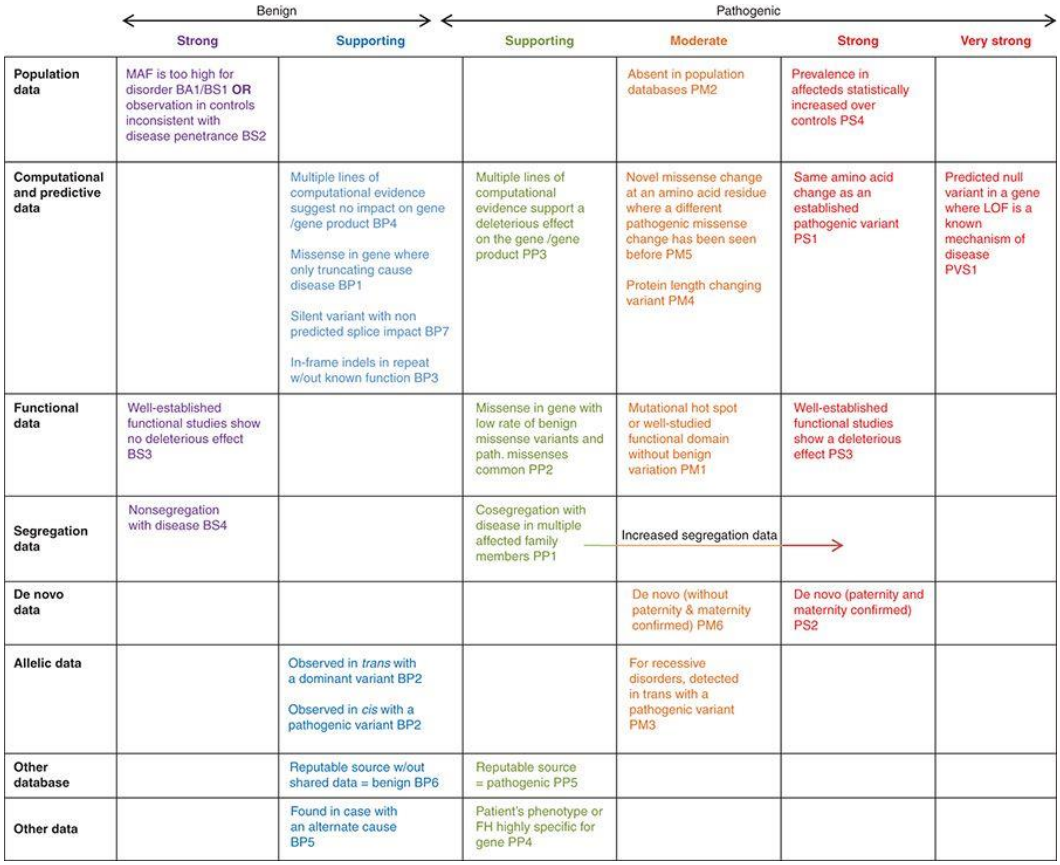


Figure 1: Evidence framework for each of the criteria for a benign or pathogenic assertion of a variant – taken from Richards et al., 2015. BS, benign strong; BP, benign supporting; FH, family history; LOF, loss of function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong.

Table 2: Rules for combining criteria to classify sequence variants – taken from Richards et al., 2015. Evidence descriptors (found in Figure 1): BS, benign strong; BP, benign supporting; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong.

Pathogenic	(i) 1 Very strong (PVS1) *AND* (a) ≥1 Strong (PS1–PS4) *OR* (b) ≥2 Moderate (PM1–PM6) *OR* (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) *OR* (d) ≥2 Supporting (PP1–PP5) (ii) ≥2 Strong (PS1–PS4) *OR* (iii) 1 Strong (PS1–PS4) *AND* (a)≥3 Moderate (PM1–PM6) *OR* (b)2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR* (c)1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Likely pathogenic	(i) 1 Very strong (PVS1) *AND* 1 moderate (PM1–PM6) *OR* (ii) 1 Strong (PS1–PS4) *AND* 1–2 moderate (PM1–PM6) *OR* (iii) 1 Strong (PS1–PS4) *AND* ≥2 supporting (PP1–PP5) *OR* (iv) ≥3 Moderate (PM1–PM6) *OR* (v) 2 Moderate (PM1–PM6) *AND* ≥2 supporting (PP1–PP5) *OR* (vi) 1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Benign	(i) 1 Stand-alone (BA1) *OR* (ii) ≥2 Strong (BS1–BS4)
Likely benign	(i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) *OR* (ii) ≥2 Supporting (BP1–BP7)
Uncertain significance	(i) Other criteria shown above are not met OR (ii) the criteria for benign and pathogenic are contradictory

1.1.10 Association studies in common complex disease

1.1.10.1 Genotyping arrays

Risk of common complex disease is contributed to by both rare and common genetic variation, although the majority of findings have been common variants (which also explain the majority of the discovered contribution to phenotypic variation) (Gibson, 2012). Genotyping arrays survey an individual's common genetic variation (MAF>1%) by determining the genotypes of between 200,000 and 2,000,000 single nucleotide polymorphisms (SNPs) distributed across an individual's genome size (Visscher et al., 2017). Genome-wide association studies (GWAS) then test whether the variation in genotype at each SNP is associated with disease status. Controls must be carefully selected from a similar population to cases (taking into consideration age, sex and possibly other covariates) (Zondervan & Cardon, 2007).

Imputation can be used to predict genotypes of SNPs not directly typed – a reference dataset of densely genotyped individuals is used to identify haplotype blocks shared between study individuals and the reference set (Marchini & Howie, 2010). Imputation is unreliable in predicting rare genotypes. Alternative methods of genotyping that can capture rare variation exist, but they each have limitations: specialised exome chips are able to genotype rare variants, but these cannot capture the full array of rare variation; high-depth WGS is most reliable but is prohibitively expensive for sample numbers required; low-depth WGS is more cost-effective but has less accuracy for rare variants; exome sequencing is also more cost-effective than WGS (though still expensive compared to genotyping arrays) but is limited to exonic variation (S. Lee, Abecasis, Boehnke, & Lin, 2014).

1.1.10.2 Association testing

In case-control studies SNP genotypes are tested for association with the phenotype using logistic regression (Al Olama et al., 2014; Nikpay et al., 2015; Wellcome, Case, & Consortium, 2007), which is able to adjust for potential confounding variables such as ethnicity, biometric information, genotyping batch or genotypes at other SNPs (Equation 1).

Equation 1: Logistic regression commonly employed in genome-wide association testing. p is the expected value of the phenotype (case = 1, control = 0), given genotype X and confounders C, D, etc. P-values are calculated based on whether β_1 significantly differs from zero.

$$logit(p) = \log\frac{p}{1-p} = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 D \ldots$$

Due to the vast numbers of SNPs tested, P-values must be corrected for multiple testing. Genome-wide significance is often stated as $P < 5 \times 10^{-8}$, the Bonferroni-corrected family-wise error rate of 0.05 considering roughly one million independent SNPs tested (M. X. Li, Yeung, Cherny, & Sham, 2012).

1.1.10.3 Power to detect association

For a variant with a particular minor allele frequency and expected effect size, there is a requirement on the number of individuals to be genotyped to achieve the statistical power to detect a genome-wide significant association (Figure 2) (Visscher et al., 2017). Both lower effect sizes and lower minor allele frequencies impose extensive and impactful requirements on the numbers of individuals required, which must be considered when designing GWAS.

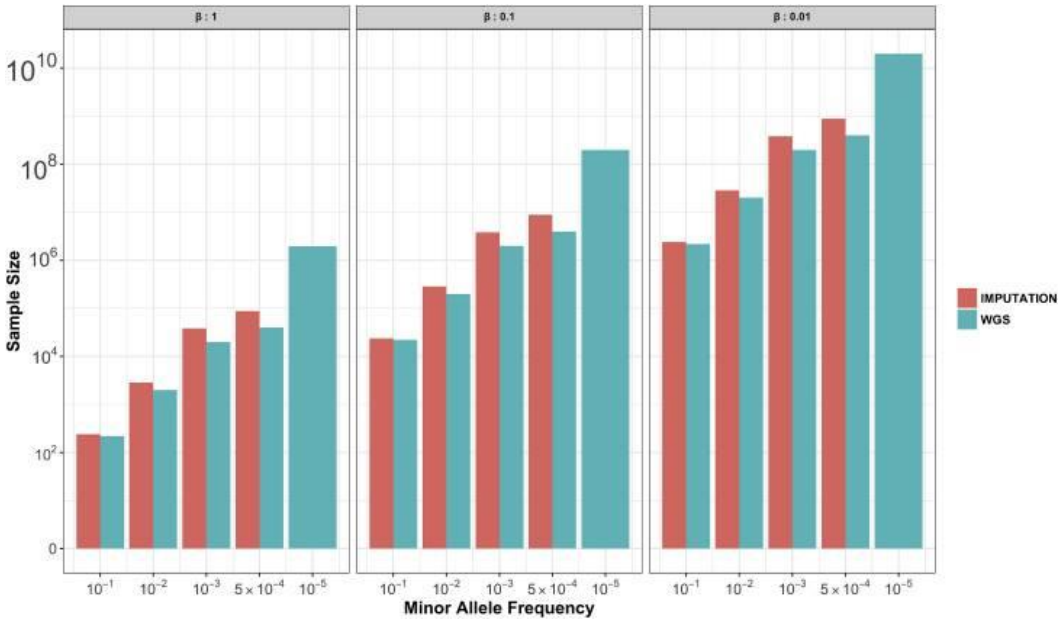


Figure 2: Minimum sample sizes for detecting trait-SNP associations from imputed and WGS data – taken from Visscher *et al*., 2017. Required sample sizes for detecting association were calculated under the assumption of a type I error rate of 5×10^{-8}, 80% power, and Hardy-Weinberg equilibrium. Effect sizes (β) are in phenotypic standard deviation units. For genotyped SNPs imputed to a fully sequenced reference, the average imputation R^2_{imp} values reported by the Haplotype Reference Consortium (McCarthy; A reference panel of 64,976 haplotypes for genotype imputation.) were used.

1.1.11 Similar phenotypes

For both rare and common complex disease, it is a key feature of genetic studies that individuals sharing the same phenotype must be identified. This may be the recruitment of multiple individuals to one study, or the identification of a known phenotype-gene relationship within an individual which has previously been discovered by genetic study.

For rare disease exome sequencing studies, individuals presenting with the same phenotypic feature or constellation phenotypic features characteristic of a particular clinical diagnosis are identified as cases (Ng, Buckingham, et al., 2010; Simpson et al., 2011). This is indicative of an implicit measure of phenotype similarity, particularly in studies of disorders where different members of the cohort have the same clinical diagnosis which can manifest as a range of different

constellations of phenotypic features (Ng, Bigham, et al., 2010). In studies where individuals are diagnosed through the identification of a known phenotype-gene relationship in the patient, phenotypes of candidate genes are evaluated with respect to the patient phenotype (The Deciphering Developmental Disorders Study, 2014; Wright et al., 2015), again suggesting implicit measures of phenotype similarity.

For common complex disease studies, there are often several criteria individuals must meet for inclusion in the study, which include manifestation of the phenotype of interest (which may include several subphenotypes), as well as age, disease severity score cut-offs, certain medical histories or family history of the condition (Al Olama et al., 2014; Nikpay et al., 2015). This is further complicated in meta-analysis datasets, where different participating members of consortia may use slightly different qualification criteria. The need to identify vast numbers of individuals to be defined as cases may result in an introduction of phenotypic heterogeneity which can reduce power to detect phenotype-specific effects - analysing more homogenous populations of cases expressing sub-phenotypes can improve the efficiency of GWAS analyses (Eichler et al., 2010; Kulminski et al., 2016; MacRae & Vasan, 2011).

1.1.12 Phenotypic data capture

The increasing abundance of available sequence data combined with increasing availability of electronic health records provides novel opportunities for genetic analyses. Considering how phenotypic information is used to define individuals with a particular disease for genetic study, it is important to capture and store these data in appropriate “machine-readable” language. It is also important to document discoveries from genetic studies in machine-readable language so that information

from individuals can be queried to this data. Phenotypic information can be recorded in several ways:

Free-text

Often, both patient clinical records and publications of novel phenotypes and disease genes are recorded in free text. Although this is the easiest and most accessible way to record observations, free-text alone renders systematic comparison across patients or between patients and known phenotypes untenable (P. N. Robinson & Mundlos, 2010). Simple text mining can be used extract meaning from free-text but to accurately translate free-text phenotype descriptions into machine-readable language sophisticated text mining and natural language processing tools are required.

Questionnaires

Phenotypic information can also be recorded in specialised questionnaires designed to gather information surrounding the particular phenotypic area of study – these can be filled by clinicians, nurses or be self-reporting questionnaires. Data recorded in this format is useful when setting inclusion requirements for cases in genetic studies of many individuals (i.e. common disease studies), allowing for the rapid identification of individuals matching specified criteria. It also enables the recording of quantitative measurements alongside binary observations of presence or absence of a phenotypic features. The use of questionnaires rather than free-text prompts the user to consider a prescribed set of phenotypes, and therefore require thoughtful design to capture meaningful information. Compliance of individuals completing questionnaires is an issue which is also relevant to the questionnaire design: questionnaires must be simple to use whilst facilitating the recording of maximal useful phenotype information.

Although questionnaires offer standardised phenotypic information to be accessed across single genetic studies, they may not facilitate comparison across studies that utilise different questionnaire designs, which presents difficulties in meta-analysis of datasets where phenotypic models are not standardised.

Controlled vocabularies

Standardised controlled vocabularies exist for the description of disease phenotypes. Closed source vocabularies for particular disease areas suffer from the same drawbacks as questionnaire usage because they do not facilitate comparison to external datasets that use different schema to record phenotypes, again making comparisons untenable (P. N. Robinson & Mundlos, 2010). However, open source vocabularies offer vast opportunities for the comparison of individual's phenotypes to other individuals or reference datasets. In extremely rare phenotypes with handfuls of cases worldwide, the power to detect disease-causing variants depends on the ability to remotely identify patients with similar phenotypes, which is likely to depend on use of a common standardised phenotypic language (as well as a data sharing strategy to facilitate). Open-source vocabularies can be continually improved by curators (although version changes can incur significant costs to widely used vocabularies (Topaz, Shafran-Topaz, & Bowles, 2013)) and external applications can be designed for the storage or analysis of phenotypic information.

A drawback to the use of controlled vocabularies is differing levels of diligence with which standardised terms are recorded. This is heavily dependent on the curation of the vocabulary and other associated tools that can facilitate easy use of the vocabulary.

1.2 Controlled phenotypic vocabulary

1.2.1 Ontology

A useful machine-readable phenotype language needs to overcome *synonymy* (many ways to express the same concept) and *hyponymy/hypernymy* (a hyponym's semantic field is more specific than its hypernym). Illustrating the issue of synonymy in phenotyping is that 'generalized amyotrophy', 'generalized muscular atrophy', and 'muscular atrophy, generalized' are all used in OMIM to describe a single entity (Amberger et al., 2014; P. N. Robinson & Mundlos, 2010). Although humans can recognise this, search engines are not well-equipped to do so. An accepted definition of the aforementioned phenotypic concept must be established, as well as all possible synonyms (which include the three examples given). Hyponymy and hypernymy are illustrated by the term 'intellectual disability' (the hypernym) encompassing 'severe intellectual disability', 'mild intellectual disability' and 'moderate intellectual disability' (hyponyms). It is useful to machine-readable phenotype language to reflect these relationships because although the hyponyms are distinct terms, they possess a high degree of similarity in their meaning. Such relationships exist throughout the entirety of medical language and are not confined to only the most similar of terms.

Ontologies are able to organise clinical definitions in a way that resolves synonymy, hyponymy and hypernymy, such that every term within the ontology represents a unique concept with defined synonyms, as well as possessing a relationship with every other term. Ontologies are described as 'a set of logical axioms designed to account for the intended meaning of a vocabulary' (Guarino, 1998) and are used in virtually every field of study to limit language complexity and organise terminologies into explicit information structures. Ontologies are

largely organised into directed acyclic graphs (DAGs), consisting of nodes and edges. In *directed* graphs edges are unidirectional, and *acyclic* graphs do not contain any paths leading from one node back to itself. Nodes nearer the root node are more general than nodes further away and ontologies often adhere to the *true-path rule* which states that all nodes assume the properties of their ancestor nodes.

1.2.2 HPO

The organisation of all human phenotypic terms into an ontology is realised in the Human Phenotype Ontology (HPO) (Köhler et al., 2014). The HPO "aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease" and contains 13,348 phenotypic terms (as of 12/06/18). In the HPO (Figure 3) each node represents a single phenotypic concept (which have a range of defined synonyms) and as with most other ontologies, terms are organised in a DAG. The root term is HP:0000001 (All) and it has 5 child terms;

- HP:0000005 – Mode of inheritance
- HP:0040279 – Frequency
- HP:0012823 – Clinical modifier
- HP:0000118 – Phenotypic abnormality
- HP:0031797 – Clinical course

'Phenotypic abnormality' subsumes the bulk of the ontology (97% of terms), containing all the phenotypic feature definitions. These can be co-annotated with terms subsumed by the other root nodes of the ontology to reflect different inheritance, onset, frequency and other clinical modifiers.

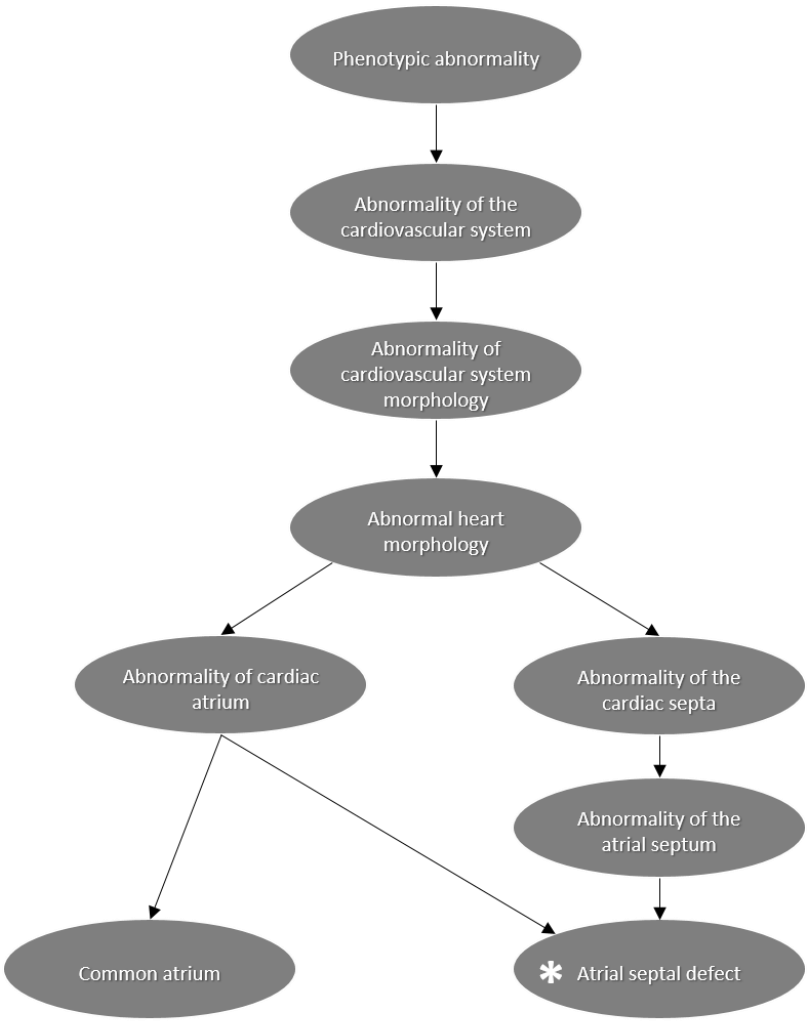


Figure 3: A subsection of the Human Phenotype Ontology. Terms are arranged in a directed acyclic graph (DAG), hence all terms are hyponyms of their parent terms and terms become more specific the further they are from the root term. Terms highlighted with '*' have multiple parent terms as they can constitute a more specific property of multiple general terms. Adapted from (P. N. Robinson & Mundlos, 2010).

The HPO was initially formed by mining for common terms across multiple diseases in OMIM Clinical Synopsis sections and making use of the hierarchical nature of these annotations to provide a primitive ontology tree structure. This ontology structure was then manually modified based on the curators' knowledge of the human genetics domain, merging similar concepts, separating generic concepts and the addition of more general terms to connect all terms to a common ancestor (Peter N. Robinson et al., 2008). The furthest level down in the ontology is 14 (i.e. distance to root is 13) and the mean distance from the terms to the root is

6.65 (Figure 4). A term's distance to the root node is correlated with a measure of its specificity (Figure 5; measured in information content (IC) which based on how few OMIM entries are annotated with a particular term, discussed later in 1.3.2). Terms further from the root tend to be more specific, but a term only 3 nodes away from the root may be considered more specific than a term over 10 nodes away – this is dependent on the construction of that particular region of the ontology and the corpus of diseases they are annotated to.

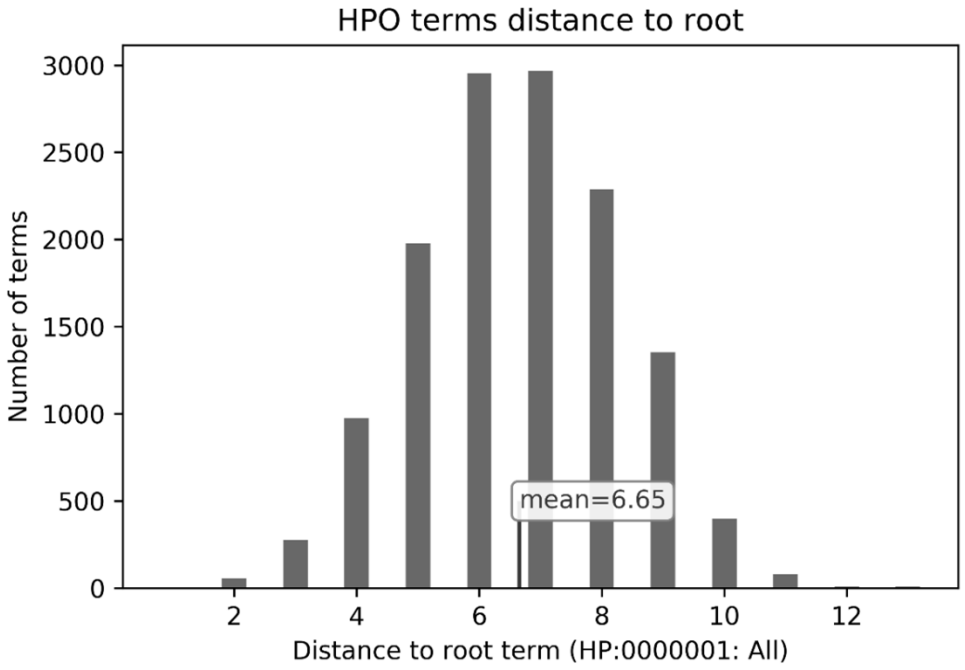


Figure 4: Distribution of distances (in number of edges) to the root HPO term (HP:0000001: All) for all nodes in the HPO (HPO build #1699, 09/02/16), showing that HPO terms most commonly exist 6-7 levels down from the root term.

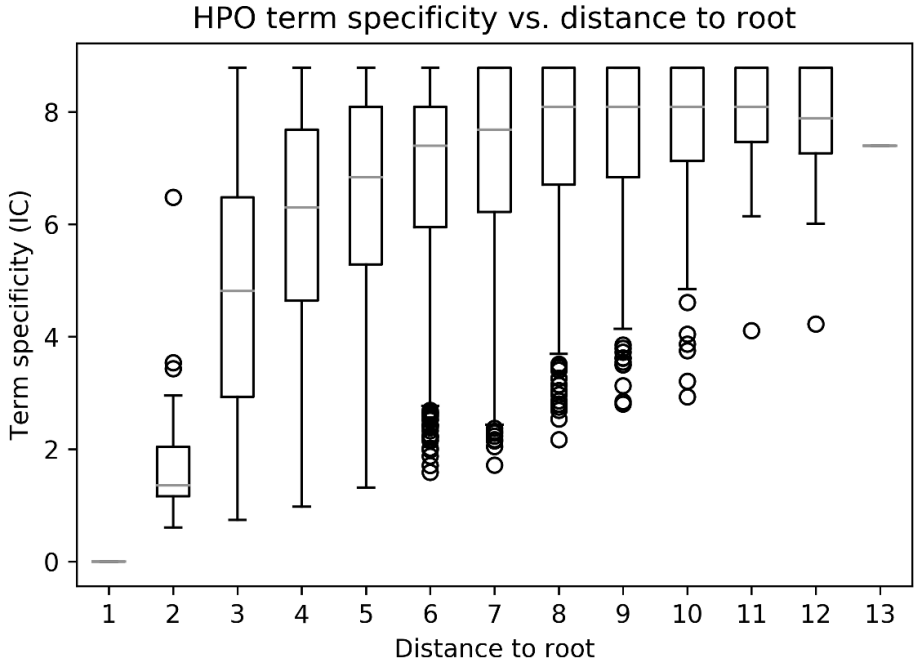


Figure 5: HPO term specificity (measured in information content: the log inverse of annotation frequency across the OMIM catalogue, discussed later in 1.3.2) vs. distance to root HPO node for all terms annotated to OMIM (n=7978; includes ancestral terms implicitly annotated). Based on the curated HPO annotation set benchmarked in Chapter 2 methodology comparisons (Table 5, page 75). Box whiskers represent furthest data points within 1.5×IQR (interquartile range).

The HPO annotations comprise 121,399 term annotations to hereditary diseases (n=8057) in OMIM, Orphanet and DECIPHER, establishing a common phenotype language for known phenotypes. The HPO can also be used to define the phenotypic characteristics of patients, as has been done in several of the largest recent rare disease sequencing projects such as the DDD (Deciphering Developmental Disorders) study (Wright et al., 2015), FORGE (Finding of Rare Disease Genes) Canada Consortium (Beaulieu et al., 2014) and the 100,000 Genomes Project (Caulfield et al., 2015). Defining disease entities (i.e a patient or a known genetic disease) with the same phenotypic language enables algorithmic comparison between them and several have been proposed (discussed in next section). The HPO also release 'negative' annotations, where HPO terms NOT presented in a particular genetic disease are listed.

1.2.3 SNOMED CT

SNOMED CT (Systematised Nomenclature of Medicine – Clinical Terms) (IHTSDO, 2014) is a standardised, multilingual clinical terminology developed over the past 40 years. It is used by physicians and healthcare providers for the storage and exchange of clinical information, as well as to support clinical decision making and analytics in software programs. Like the HPO, it comprises systematically organised computer-processable medical terms, providing codes, terms, synonyms and definitions, although it attempts to cover all medical terminology rather than only rare genetic disease phenotypes, resulting in far more terms in total (Table 3). It is reportedly used in over 50 countries (D. Lee, Cornet, Lau, & de Keizer, 2013) and countries including the United States, United Kingdom, Canada, New Zealand and Australia have designated SNOMED CT as the recommended clinical reference terminology for clinical information systems

(Australian Digital Health Agency, n.d.; Canada Health Infoway, n.d.; Ministry of Health - New Zealand Government, 2017; NHS, n.d.; U.S National Library of Medicine, 2018). There is a wealth of SNOMED CT annotations contained in electronic health records of user countries and OMIM clinical synopses also contain mappings to SNOMED (Amberger et al., 2014).

1.2.4 ICD

The International Classification of Disease (ICD) is a system for the classification of disease to permit the "recording, analysis, interpretation and comparison of mortality and morbidity data collected in different countries or areas and at different times" (World Health Organisation, 2010). It is primarily used to record hospital episode statistics so that the maximal amount of machine-readable clinical information is available on health records, as well as for the tracking of epidemiological trends and for billing purposes. OMIM disorders also contain mappings to ICD codes (Amberger et al., 2014).

Recently, clinical coding on electronic health records in the form of ICD codes has been linked with genetic data in several projects, such as HUNT (Krokstad et al., 2013), MGI (Michigan Genomics Initiative) (The Michigan Genomics Initiative, 2016), DiscovEHR (Dewey et al., 2016) and UK Biobank (UK Biobank, 2018). ICD codes have been used to recruit individuals for GWAS (Howard et al., 2018; Mitchell et al., 2016) They have also been used for PheWAS studies, where genotype data is probed for an association with a range of clinical phenotypes, commonly as ICD codes (Denny et al., 2013, 2011, 2010; Ritchie et al., 2013; Verma et al., 2018).

Table 3: Number of clinical terms contained in the HPO, SNOMED and ICD-10 ontologies (comparison undertaken in 2014).

Ontology	**Number of terms**
HPO *(all)*	~11,000
- HPO *(phenotypic abnormality)*	~10,000
SNOMED *(all)*	~300,000
- SNOMED CT *(clinical finding)*	~100,000
- SNOMED CT *(disorder)*	~60,000
ICD-10-CM *(clinical modification)*	~90,000

1.2.5 Other biomedical ontologies

The National Center for Biomedical Ontology (Musen et al., 2012) lists 716 ontologies, spanning fields such as medical procedures, drugs, anatomy and many organism-specific ontologies. Relevant biomedical ontologies to human phenotypes (P. N. Robinson & Mundlos, 2010) include Gene ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium, 2017), the Mammalian Phenotype Ontology (MPO) (Smith & Eppig, 2012), the Foundational Model of Anatomy (FMA) ontology (Rosse & Jr., 2007), the Sequence Ontology (Eilbeck et al., 2005), the Cell Ontology (Bard, Rhee, & Ashburner, 2005), the Chemical Entities of Biological Interest (ChEBI) ontology (Hastings et al., 2016) Orphanet Rare Disease Ontology (ORDO) (Maiella et al., 2013) and [human] Disease Ontology (DO) (Kibbe et al., 2015).

1.2.6 Why this thesis will focus on HPO

The HPO is a clinical resource, specifically designed for the study of rare diseases, whereas SNOMED CT and ICD are both primarily used for health provider analytics and billing and the user guide for the ICD states that “The ICD is neither intended nor suitable for indexing of distinct clinical entities” (World Health Organisation, 2010). The HPO is the dedicated ontology for the phenotyping of

patients on several rare disease projects (Beaulieu et al., 2014; Caulfield et al., 2015; The Deciphering Developmental Disorders Study, 2014), and much of the published methodology for storing and using patient phenotypes to identify differential diagnoses and causative genes is based on the HPO (Bauer, Köhler, Schulz, & Robinson, 2012; Girdea et al., 2013; Köhler et al., 2009; Peter N. Robinson et al., 2014; Smedley et al., 2014).

1.3 Established HPO methodologies

1.3.1 Calculating similarity between two sets of HPO terms

Use of standardised phenotypic language within an ontology structure to describe disease has enabled algorithmic comparison between two disease entities composed of phenotypic terms. Algorithmic scoring of similarity between sets of disease entities could be either:

- Patient to patient: for establishing cases of similar phenotypes in gene discovery (and potentially dissimilar cases to use as controls).
- Patient to known disease: for establishing most likely phenotypic matches to patient(s), enabling personalised lists of candidate genes.
- Known disease to known disease: for analysis of similarity between known diseases and to reveal common disease aetiologies and gene pathways.

There are many different methods that can be used to calculate similarity between two entities comprising ontology terms, many of which have been proposed to algorithmically calculate similarity between groups of gene ontology (GO) terms (Pesquita, Faria, Falcão, Lord, & Couto, 2009) – these methods will be discussed in the context of calculating similarity between disease entities, which are represented as constellations of phenotypic terms, regardless of whether they may have originally been developed for similarity between HPO terms, GO terms or for other semantic similarity applications. Measuring similarity between two sets of groups of ontology terms generally consists of two stages. Firstly, a measure of similarity is defined between individual terms, and secondly, these individual term-wise similarities are aggregated to measure similarity between groups of terms (diseases).

1.3.2 Term-wise similarity

Similarity between terms can be measured based on nodes or edges (or hybrid methods). Edge-based similarity commonly relies on distance between two nodes (or between the nodes and their common ancestor). However, edge-based methods are rarely used because they rely on nodes and edges being uniformly distributed throughout the ontology, and that edges within an ontology represent a consistent semantic distance – often, neither of these assumptions are true within biomedical ontologies (Pesquita et al., 2009).

Node-based measures compare terms based on properties of the terms themselves, which could relate to their ontological relationships or information external to the ontology (e.g. probability of finding them within a particular corpus). A metric that is often incorporated is information content (IC), a measure of how specific a term is, and the central idea is that a pair of terms with a specific common ancestor should be scored as more similar than a pair of terms with a more general common ancestor. The IC of a term is calculated by taking the negative log likelihood of it appearing in the corpus (Equation 2) – for the purposes of HPO phenotypic similarity, this could be the likelihood of it being annotated to a particular OMIM disorder. Note that the corpus used will bias the IC values for each term – different knowledge bases will represent different terms (or sub-groups of terms within an ontology) at differing frequencies which will affect how specific each term is perceived to be.

Equation 2: Information Content of term *t*

$$(t) = -\log p(t) = -\log\left(\frac{phenotypes\ annotated\ with\ t}{total\ phenotypes}\right)$$

Table 4: OMIM frequency and resultant information content (IC) for selected HPO terms from Figure 1. Column 2 denotes the position on the ontology in Figure 6.

HPO term	*Fig. 6*	*OMIM freq.*	*IC*
Phenotypic abnormality - HP:0000118	1	6518	0
Abnormality of the cardiovascular system - HP:0001626	2	1688	1.35
Abnormality of cardiovascular system morphology - HP:0030680	3	961	1.91
Abnormal heart morphology - HP:0001627	4	869	2.01
Abnormality of cardiac atrium - HP:0005120	5	213	3.42
Abnormality of the cardiac septa - HP:0001671	6	369	2.87
Abnormality of the atrial septum - HP:0011994	7	212	3.43
Atria septal defect - HP:0001631	8	211	3.43
Common atrium - HP:0011565	9	2	8.09

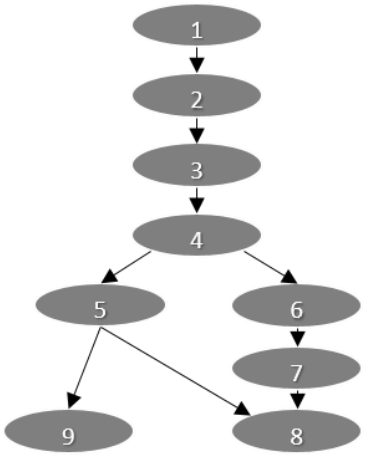


Figure 6: Simplified representation of the subsection of the HPO in Figure 3, for reference of Table 4.

Table 4 shows the IC of selected nodes within the HPO. Note that child nodes always have higher IC values than their parents due to implicit ancestor annotation. The Resnik measure (Equation 3) uses the IC of the most informative common ancestor (MICA) to calculate similarity between terms (Resnik, 1999). Using Resnik as an example, the highest similarity between terms in Table 4 would be two ‘Common atrium’ terms (IC = 8.09) as logically, the MICA of two identical terms would be the term itself. This would be followed by ‘Atria septal defect’ and ‘Common atrium’, for whom the MICA would be ‘Abnormality of the cardiac atrium’ (IC = 3.42). The similarity between ‘Abnormality of cardiac atrium’ and

'Abnormality of the cardiac septa' (or two terms which specific properties of one each of these, but not both) would be lower as the MICA is the much more general 'Abnormal heart morphology' (IC = 2.01).

Equation 3: Resnik similarity between two terms (Resnik, 1999).

$$sim(t1, t2) = \quad (MICA(t1, t2))$$

Other node-based measures including Lin (Equation 4) and Jiang-Conrath (Equation 5) use modified measures of term specificity which also consider the relative specificity of the MICA in comparison to that of the terms (therefore measuring distance within the ontology).

Equation 4: Lin similarity between two terms (Lin, 1998).

$$sim(t1, t2) = \frac{2 \times \quad (MICA(t1, t2))}{(t1) + \quad (t2)}$$

Equation 5: Jiang-Conrath similarity between two terms (Jiang & Conrath, 1997).

$$sim(t1, t2) = \quad (t1) + \quad (t2) - 2 \times \quad (MICA(t1, t2))$$

1.3.3 Aggregating term-wise similarities between two disease entities

1.3.3.1 Pairwise

With a schema that defines similarity between a pair of terms, it is then necessary to define how these similarities would be aggregated when comparing two groups of terms. Pairwise approaches involve taking an average of the pairwise combinations between terms. An average of all pairwise combinations doesn't perform well as it includes far too many terms in the equation, and instances of high similarity between particular terms lose relevance when the average is calculated. Conversely, using the maximum similarity value between all pairwise combinations of the terms excessively reduces the information. One solution is to use the best-match average (Equation 6) which balances between the two, and this

has proved to be beneficial when benchmarking phenotypic similarity performance (Buske, Girdea, et al., 2015; Pesquita et al., 2009). The best-match average gives asymmetric similarities between two annotated phenotypes [$sim(Q,D) \neq sim(D,Q)$] so similarities are often made symmetrical using Equation 7. The use of Resnik to measure similarity between term nodes (Equation 3) combined with best-match average term similarity (Equation 6) and conversion to a symmetrical score (Equation 7) will be referred to as $\text{Resnik}_{\text{avg,max|sym}}$ from hereon.

Equation 6: Best-match average term similarity between diseases Q and D

$$sim(Q \rightarrow D) = avg\left[\sum_{t1\ \epsilon\ Q} \max_{t2\ \epsilon\ D} sim(t1, t2)\right]$$

Equation 7: Symmetrical similarity between diseases Q and D.

$$sim_{symmetric}(Q, D) = \frac{1}{2} sim(Q \rightarrow D) + \frac{1}{2} sim(D \rightarrow Q)$$

1.3.3.2 Groupwise

Groupwise measures simplify the ontology graph to sets of terms so that set similarity methods can be used. One such example is the Jaccard index (Equation 8), which can be modified to weight terms by IC (Pesquita, Faria, Bastos, Falcão, & Couto, 2007). Groups of terms can also be represented as vectors, using the cosine rule to evaluate similarity (Equation 9). Vectors could be binary (set to 1 if disease is annotated HPO term; 0 if disease isn't annotated with HPO term), although vector models are frequently used in information retrieval as vector features can be weighted by information content, as well as the frequency with which terms are found in text (which may have relevance in cases where certain phenotypic features have variable prevalence in different diseases). Quantification of phenotype terms and the use of vectors and cosine similarity has been used

outside of a clinical context in human phenome analysis (Lage et al., 2007; van Driel, Bruggeman, Vriend, Brunner, & Leunissen, 2006).

Equation 8: Jaccard index, measuring similarity based on the size of the intersection between two groups of terms divided by the size of the union.

$$J(Q, D) = \frac{|Q \cap D|}{|Q \cup D|}$$

Equation 9: Cosine similarity between vectors Q and D

$$sim(Q, D) = \cos\theta = \frac{Q \cdot D}{\|Q\| \cdot \|D\|}$$

1.3.4 Differential diagnosis

There are existing tools that calculate similarity between a group of HPO terms describing a patient and the full knowledge-base of human disease. One such example is the **Phenomizer** (Köhler et al., 2009), which uses a best-match $Resnik_{avg,max|sym}$ followed by an estimation of statistical significance to provide differential diagnostic phenotypes. **BOQA** utilises a groupwise graph-based approach, weighting similarity between nodes based on the probability of type I or II error associated with observing the similarities and differences of a query set of terms and a reference set of terms.

1.3.5 Phenotype data storage

Phenotips (Girdea et al., 2013) is a an open source software for the collection and analysis of patient phenotypic information, primarily in the format of HPO terms. It also facilitates data entry for patient demographics, medical history, family history, physical and laboratory measurements, physical findings and additional notes. Phenotips has been implemented as the standard patient information record in large rare disease sequencing projects such as FORGE/CARE4RARE (Beaulieu

et al., 2014) and the NIH Undiagnosed Diseases Program (UDP) (Gahl et al., 2016), as well as in genetics clinics in the UK/worldwide.

Phenotips contains an OMIM phenotype recommender based on semantic similarity metrics like those of the Phenomizer discussed previously. However, this implementation also accounts for general population frequency of disorders taken from Orphanet (Rath et al., 2012) (i.e. more frequent disorders in general population more likely to be recommended), negative phenotypes, and handles free-form text inputs. However, this hasn't been benchmarked as a diagnostic tool and exists only to estimate and recommend a list of the 20 most likely disorders. Phenotips has also integrated Phenomizer (Köhler et al., 2009) and BOQA (Bauer et al., 2012) similarity calculations for patients.

1.3.6 Phenotype-driven genome/exome interpretation

Several tools have been developed that incorporate disease similarity calculations for the prioritisation of variants.

EXtasy (Sifrim et al., 2013) combines a number of different variant annotations (which consist of conservation and deleteriousness scores) with a gene phenotype score, which is generated by calculating similarity between the candidate gene and known genes associated with the patient phenotype, based on shared gene annotations using the Endeavour algorithm (Aerts et al., 2006). A random forest classifier is used to combine the variant annotations and the gene phenotype score into a final variant score.

Phevor (Phenotype Driven Variant Ontological Re-ranking) (Singleton et al., 2014) makes use of the annotated links between the HPO (Human Phenotype Ontology), MPO (Mammalian Phenotype Ontology), DO (Disease Ontology) and

GO (Gene Ontology), propagating information about patient phenotypes (encoded in HPO terms) across and throughout these ontologies to build candidate gene scores. Candidate gene scores are then combined with variant scores to produce a final prioritisation score.

Phen-Gen (Javed, Agrawal, & Ng, 2014) scores both coding and non-coding variants in a unifying framework that estimates their highest possible impact, using different data sources for different types of mutation. Gene phenotype scores are calculated using Phenomizer (Köhler et al., 2009) probabilities, and these gene scores are then expanded to genes not previously associated with disease using a random walk algorithm across a gene-gene interaction network. For each gene, the variant and gene phenotype scores are then combined using a Bayesian framework.

PhenIX (Zemojtel et al., 2014) combines variant frequency and deleteriousness prediction tools to score exome variants. Phenomizer (Köhler et al., 2009) similarity metrics are used to score genes by their disease phenotype similarity to the patient's HPO terms. Only genes mapped to Mendelian phenotypes from OMIM/Orphanet/DECIPHER databases (termed the disease-associated genome) are scored with respect to phenotype. The variant and gene phenotype scores are combined into a single score for variant prioritisation, including a step where the variant score is modified based on the suspected mode of inheritance.

PHIVE (Peter N. Robinson et al., 2014) uses the PhenoDigm algorithm (a variant of the best-match Resnik algorithm with a final normalisation step) (Smedley et al., 2013) to calculate similarity between patient HPO terms and mouse mutants associated with each gene, making use of mappings between the HPO and the Mammalian Phenotype Ontology (MPO). Although a finding in a gene previously

unreported to be involved in human disease may be insufficient to issue a clinical diagnostics report, utilising mouse models enables a much higher coverage of the protein-coding exome for genetic analysis and can provide good candidates for follow-up.

hiPHIVE (Smedley et al., 2015) incorporates the mouse models of PHIVE, as well as zebrafish and human phenotype data and protein-protein interaction data.

1.3.7 Phenotype matchmakers

As mentioned in 1.1.6, the discovery of novel genes in rare monogenic disorders requires recruitment of sufficient patients with the same disease to perform genetic analysis with the ability to rule out non-pathogenic variation (hence, honing in on pathogenic variation). Disease cases can often be both rare and sparsely located across the world – usually not united by a single healthcare record system. There are several phenotype "matchmakers" employed worldwide that link rare disease patient phenotype data to genetic data, and utilise HPO-based methods to identify similar patients so that genetic analyses can be undertaken. DECIPHER (Chatzimichali et al., 2015), GeneMatcher (Sobreira, Schiettecatte, Valle, & Hamosh, 2015) and PhenomeCentral (Buske, Girdea, et al., 2015) all facilitate the storage of patient phenotype and genetic data, each offering internal phenotype and gene matchmaking services to identify similar patients using HPO ontology similarity metrics previously discussed. These matchmakers are also united to enable external matching across the three different platforms with the Matchmaker Exchange API, which enables patients external to the aforementioned databases to be queried to each database (Buske, Schiettecatte, et al., 2015).

1.4 Challenges of phenotype similarity methodology addressed in this thesis

1.4.1 Drawbacks of existing phenotype similarity and diagnostic methodologies

Existing candidate variant prioritisation methods have been shown to be effective in a handful of cases or simulated patients/exomes (Peter N. Robinson et al., 2014; Singleton et al., 2014; Zemojtel et al., 2014), although the underpinning phenotype similarity metrics have limitations. Clinical features are annotated to phenotypes as binary present/absent observations which are unable to describe the relevance of each phenotypic feature to the overall disease. For example, primary microcephaly-1 (MIM #251200) is characterised by 16 'Phenotypic abnormality' terms in the curated HPO annotation set, including the core feature 'Microcephaly' (HP: 0000252) as well as other features of lower penetrance (such as 'Renal hypoplasia/aplasia' – HP: 0008678). Binary annotation is unable to reflect the relative importance of terms in similarity calculations, in this case weighting the cardinal 'Microcephaly' feature equally to non-obligate features such as 'Renal hypoplasia' and 'Hyperreflexia'. Although penetrance data is recorded for a proportion of HPO phenotype annotations (41% of current HPO curated annotations have quantification information with 48% of diseases containing at least one quantified phenotype term), pairwise term similarity measures such as the $Resnik_{avg,max|sym}$ do not utilise this information, although BOQA (Bauer et al., 2012) is a Bayesian query tool that has been built to utilise the limited existing penetrance information. Regardless, it is unclear how non-quantified terms, which remain in the majority, should be encoded and whether they should be assumed to be fully penetrant to the disease or not – it is uncertain if the annotations without associated

frequency information belong to this distribution of recorded annotation frequencies (Figure 7) or can be assumed to be fully penetrant.

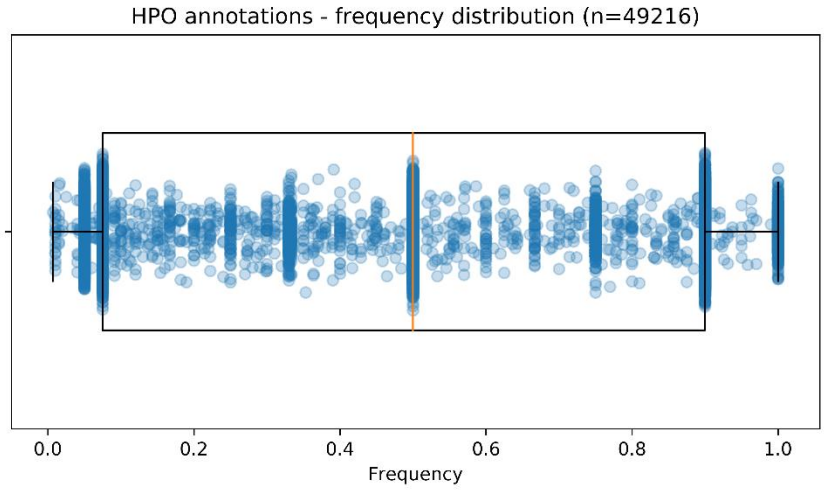


Figure 7: Boxplot showing the wide range of penetrance frequencies associated with annotations of HPO terms to OMIM diseases (n=49,216 annotations), including conversion of text-based frequencies ("Hallmark", "common", etc) according to guidelines (P. N. Robinson & Mundlos, 2010). Based on curated OMIM HPO annotation set benchmarked in Chapter 2 used for metrics (Table 5, page 75). Box whiskers represent furthest data points within 1.5 × IQR (interquartile range).

Considering these limitations, this thesis investigates the use of a quantitatively annotated reference disease set, where HPO terms were weighted by relevance to their diseases (Figure 8). There is a wealth of free-text describing disease phenotypes, particularly in OMIM where it is already directly mapped to the disease (and genes). Simple text mining of these free-text disease descriptions is able to provide an effective weighting schema that can estimate the relative importance of each phenotypic characteristic to each phenotype, and this was utilised to generate phenotype annotations, using term frequency to approximate relevance.

The utility of text-mined reference annotations was established by comparison to the curated annotation set of the same diseases employed by HPO-based tools that perform phenotype similarity searches between queries and reference human

Mendelian disorders (Bauer et al., 2012; Köhler et al., 2009). The quantified text-mined phenotype annotations were also compared against an unquantified version of the same annotation set. Secondly, a vector space model (VSM) was used that evaluates cosine similarity between HPO-annotated diseases, comparing this to $Resnik_{avg,max|sym}$ similarity implemented in the Phenomizer, as well as the BOQA algorithm (Bauer et al., 2012). In this thesis it is hypothesised that the use of text-mined phenotype term quantification to represent relevance, combined with a suitable similarity measure, would enhance our ability to identify similar diseases.

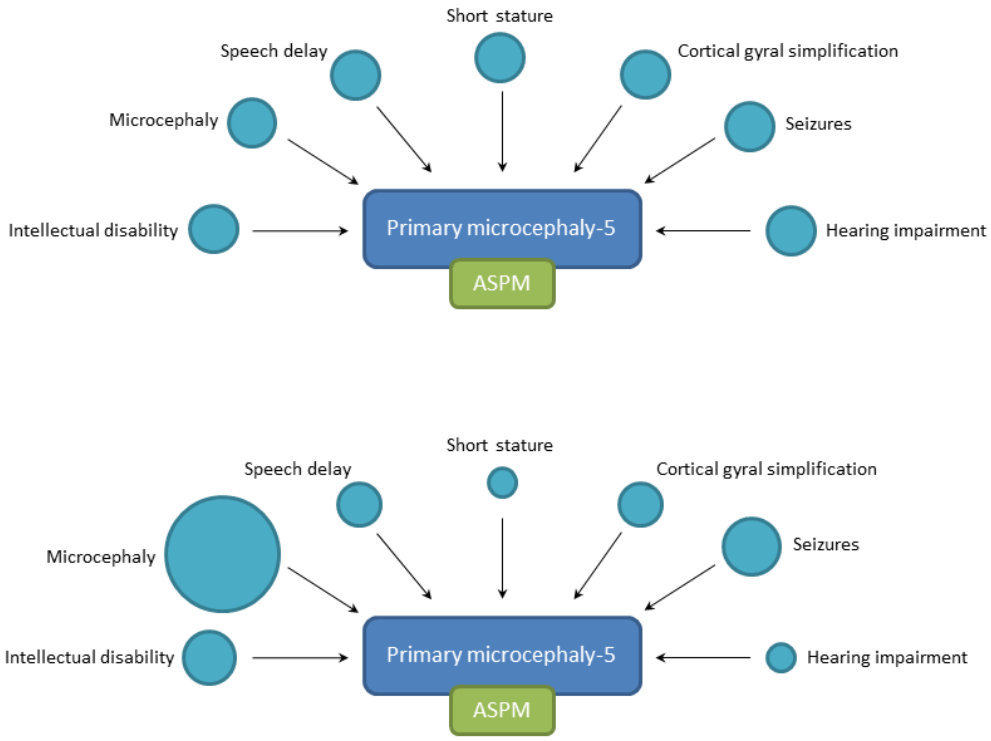


Figure 8: Schematic representing how text-mining can quantify the relevance of certain phenotypic characteristics to the overall disease (lower panel) compared to a lack of quantification (upper panel), using Primary microcephaly-5 (OMIM:605481) caused by ASPM gene as an example. In the lower panel, frequency of HPO terms found in the OMIM description is roughly represented by the diameter of the circles.

Benchmarking procedures employed in this thesis range from simulation of phenotype queries using known clinically similar entries on OMIM (Phenotypic Series) (Amberger et al., 2014), through to testing on real patient data. Phenotype

similarity methodologies were used to firstly to predict DDD patient diagnoses based on their clinician-assigned HPO terms (Wright et al., 2015), and then used to predict causative variants in patients from the Guy's Hospital genetics clinic which had also been assigned HPO terms (n=103).

1.4.2 Common complex disease phenotypes from questionnaire data

In common complex disease, collected phenotype data can be used for analysis of subphenotypes (Nikpay et al., 2015) or extreme phenotypes (Al Olama et al., 2014), though these are often limited to the selection of one or few subphenotypes. There is an increasing number of studies using collected phenotype data from electronic health records to conduct genome-wide association analysis on common disease, which utilise coded data (e.g. ICD9-10-11) (Dewey et al., 2016; Krokstad et al., 2013; The Michigan Genomics Initiative, 2016; UK Biobank, 2018), some of which is self-reported. Here, extensive phenotype questionnaire data collected in severe acne patients is used to systematically identify subphenotypes for which there is power to conduct novel GWAS analyses. The analysis of more homogenous populations of cases expressing sub-phenotypes can empower GWAS studies to identify novel loci associated with specific disease subphenotypes (Eichler et al., 2010; Kulminski et al., 2016; MacRae & Vasan, 2011), and it is hypothesised that the increased granularity of phenotype information recorded in questionnaires will enable novel associations to be identified between genomic loci and subphenotypes of acne.

Chapter 2 - Improvements in measures of rare disease phenotype similarity

2.1 Introduction

Documented human phenotype-gene pairs have been curated for the purpose of cataloguing genes, genetic phenotypes and the ever-increasing numbers of reported gene-phenotype causal relationships, to enable user-friendly searching for clinical and molecular genetic researchers (Amberger et al., 2014). The representation of phenotypes in free-text is problematic for standard searching algorithms, but the majority (84% ; 6,902 of 8,184 in February 2016) of OMIM phenotypes are now also represented as composites of HPO terms – a machine-readable language that is more suited to algorithmic searching (P. N. Robinson & Mundlos, 2010). Despite the initial automated concept recognition stage used during the inception of the HPO to create the first set of OMIM-HPO annotations, these annotations are manually curated by the developers of the HPO and any medical concept initially annotated automatically to an OMIM disorder was later manually ratified (Peter N. Robinson et al., 2008).

Querying patient phenotype information to known phenotype-gene pairs is used for the suggestion of differential clinical diagnoses for patients, as well as for the identification of candidate genes for genetic diagnoses (which can be found through mapping to genes of similar phenotypes with known genetic cause). Many similarity metrics have been proposed for the calculation of similarity between two disease entities comprising HPO terms, including:

- Symmetrical, best-match average Resnik ($Resnik_{avg,max|sym}$) (Resnik, 1999) similarity measure, which is the basis of the Phenomizer (Köhler et al., 2009), PhenIX (Zemojtel et al., 2014) and Phen-Gen (Javed et al., 2014).
- BOQA, a Bayesian graph-based measure weighted by probability of type I and II errors involved in differences between reference disease annotation and query.
- PhenoDigm, used to calculate similarity between human phenotypes and mouse models (Smedley et al., 2013), the basis of PHIVE exome variant prioritisation (Peter N. Robinson et al., 2014).

The phenotype similarity methods that calculate patient similarity to a human disease reference set utilise the curated HPO phenotype annotations to characterise reference diseases. Monogenic disease phenotypes often consist of 'core' features (presented by every patient with the disease) as well as 'non-obligate' features (not presented by every patient with the disease, with variable penetrance), and it is important that these are represented in annotations and similarity modelling. The HPO annotations contain such frequency information, but it doesn't cover all diseases (41% of annotations have associated penetrance data, with 48% of diseases having one or more quantified term). However, pairwise term similarity measures do not account for these frequencies. Only the graph-based BOQA is able to account for these frequencies but it is still uncertain how the majority of terms that do not possess associated frequency information should be encoded, especially considering the frequency distribution of known penetrance information in the HPO (Figure 7).

Here, the use of a text-mined reference set of phenotypes was investigated in comparison to the manually curated set, using the frequency with which terms

appear in free-text phenotypic descriptions as a simple weighting schema that encompasses all phenotypes. Also, the use of feature vectors and cosine similarity were investigated as a similarity metric, which was hypothesised to perform more effectively than a pairwise similarity metric in representing quantified phenotypes. Cosine similarity was compared to $Resnik_{avg,max|sym}$, the basis of a clinical diagnostic disease similarity tool and several variant prioritisation tools. The performance of these phenotype matching algorithms was benchmarked using OMIM phenotypes judged to be clinically similar (in the same OMIM phenotypic series), removing them from the reference set to be used as a query phenotype.

2.2 Materials and Methods

2.2.1 Phenotype annotation

2.2.1.1 Curated OMIM phenotype annotations

For comparison with text-mined OMIM phenotype annotations, the publicly available manually curated OMIM phenotype annotations were used (build #1233, Jan 13 2016, downloaded 09/02/16). Only 'phenotypic abnormality' annotations were used to ensure equivalence with annotations used by Phenomizer.

2.2.1.2 Text-mined phenotype annotation

Text-mined phenotype annotations were obtained by extracting phenotype free-text descriptions from OMIM (date: 05/02/16) and submitting to the NCBO Annotator API (Shah et al., 2009) to identify instances of HPO terms. Annotator is an open Web service made available by the National Centre of Biomedical Ontology (Musen et al., 2012) which infers ontology annotations from text. Annotator utilises 'an exact string comparison (a "direct" match) between the text and ontology term names, synonyms, and IDs' (Shah et al., 2009). Again, HPO terms were filtered to include only 'Phenotypic abnormality' terms (HP: 0000118) to ensure equivalence to the manually curated OMIM phenotype annotations.

Due to the different numbers of annotated OMIM phenotypes available for the curated set and text-mining (6,902 and 7,600 respectively), benchmarking was performed on the intersection of these phenotypes (n=6,518 –Table 5, page 75).

2.2.2 Phenotype similarity calculation

2.2.2.1 $Resnik_{avg,max|sym}$ (Equation 3, Equation 6, Equation 7)

Once annotated, similarity between phenotypes was calculated. The performance of the symmetrical max-average Resnik algorithm implemented in the Phenomizer

(Köhler et al., 2009; Resnik, 1999) was compared to cosine similarity, which is used in various information retrieval methods.

Equation 3 (repeat): Resnik similarity between two terms (Resnik, 1999).

$$sim(t1, t2) = \quad (MICA(t1, t2))$$

Equation 6 (repeat): Best-match average term similarity between diseases Q and D

$$sim(Q \rightarrow D) = avg\left[\sum_{t1 \in Q} \max_{t2 \in D} sim(t1, t2)\right]$$

Equation 7 (repeat): Symmetrical similarity between diseases Q and D.

$$sim_{symmetric}(Q, D) = \frac{1}{2} sim(Q \rightarrow D) + \frac{1}{2} sim(D \rightarrow Q)$$

2.2.2.2 Cosine similarity (Equation 9)

Diseases are represented as feature vectors, with each dimension indicating the presence (or absence) of a particular HPO term. Similarity between two disease vectors is measured using the cosine of the angle between them. In this application of cosine similarity, the score ranges from 1 (corresponding to an angle of 0^0, indicating identical vectors) to 0 (corresponding to 90^0, indicating orthogonality).

In its simplest form a vector space model requires vector components to be set to 1 (indicating the HPO term is present) or 0 (indicating the HPO term is absent). It can be advanced from this simple binary setting using a variety of techniques (this implementation incorporates all of the following):

Use of the semantic inheritance structure of terms; recalling that annotation of a particular ontology term implicitly annotates every ancestor of that term on the DAG, ancestral terms can also be included in the disease vectors.

Using term frequency; pertinent when using text-mined annotations, the number of times a disease is annotated with a particular HPO term is used.

Use of term weights; vector components can be modified by multiplying them by the information content (IC) of their terms (Equation 2). This up-weights vector features that correspond to specific terms.

Equation 9 (repeat): Cosine similarity between vectors Q and D

$$sim(Q, D) = \cos\theta = \frac{Q \cdot D}{\|Q\| \cdot \|D\|}$$

2.2.3 Simulated queries using OMIM Phenotypic Series

When benchmarking information retrieval methods, approaches generally involve defining a set of entities (diseases) within the corpus that are 'known' to be similar (by consulting literature/experts or generating simulations) and then observing how well a particular method performs in classifying such entities as similar. The approach developed here made use of the OMIM phenotypic series (PS) as the set of known similar diseases. The OMIM phenotypic series are defined as a set of 'phenotype entries [that] overlap significantly in their clinical manifestations', and that they are classified according to clinical judgement, not computed similarity (Amberger et al., 2014). Phenotypic series have variable size, grouping from 2 to 78 diseases (mean = 8.03). There are 353 phenotypic series, covering 2785 diseases in the OMIM catalogue (PS information downloaded: 15/02/16).

For each disease in a phenotypic series, the disease was removed and its terms used as a query to the remaining OMIM reference set. The diseases in the remaining set were ranked by similarity to the query, evaluating based on the ranks of the diseases within the same phenotypic series. Methods were evaluated by conducting this test on all diseases in phenotypic series and aggregating the results. Only diseases that were annotated by all methods (subscript *N* in Table 5, page 75) were included in

the analysis and diseases in multiple phenotypic series were not used (n=47), leaving 2,317 phenotypes.

2.2.4 Evaluation of simulated queries

The performance of an individual simulated query is evaluated using the list of similar phenotypes generated (n=6,517, one less than the intersection of all phenotypes), and based on where the phenotypes in the same PS are situated in this list – the higher they are ranked, the better the performance of the method. Benchmarking in this chapter does not consider the magnitude of phenotype similarity score, only the rankings of phenotypes by similarity score. Due to the variable size of the OMIM PS, 'Top N' measures of sensitivity are not appropriate, nor is only presenting the distributions of the ranks and performing simple statistical hypothesis tests to compare between methods.

2.2.4.1 Interpolated precision-recall

Receiver operating characteristic (ROC) metrics are commonly used to evaluate binary classifier models (the results of the query are dichotomised into 'similar' and 'not similar'). The measures that contribute to the ROC curve are true positive rate (recall/sensitivity; Equation 10) and false positive rate (Equation 11), calculated at a series of ranks and plotting the two measures against each other to form a curve, the area under which is descriptive of the performance of the model.

Equation 10: Recall (*R*) for a query at rank *r*. Also known as sensitivity or true positive rate.

$$R_r = \frac{similar\ diseases\ with\ \ \ rank\ r}{total\ similar\ diseases}$$

Equation 11: False positive rate (*FPR*) for a query at rank *r*.

$$FPR_r = \frac{nonsimila\ \ diseases\ with\ \ \ rank\ r}{total\ nonsimilar\ diseases}$$

Due to the skewed class distribution of this benchmarking (a maximum of 2% of the query results are classed “positive”; the rest will be classed “negative”), ROC metrics are unsuitable as the relatively large number of “negatives” (the denominator of Equation 11) will keep the FPR very low and unable to distinguish between different methods. Precision is often used instead of FPR in information retrieval evaluation where the class distribution is skewed (Fawcett, 2006). Starting from rank 1 and iterating through further ranks, the precision and recall were calculated (defined in Equation 12 and Equation 10 respectively) for each query.

Equation 12: Precision (*P*) for a query at rank *r*.

$$P_r = \frac{similar\ diseases\ with\quad rank\ r}{r}$$

To overcome difficulties in averaging performance over queries with variable numbers of positive results, an interpolation step was included (Manning & Raghavan, 2009). This involved defining 11 standard recall points (0 … 0.1 … 0.2 … … 1) and using the maximum precision found above each point as the 11-point interpolated precision recall (Equation 13).

Equation 13: Interpolated precision (*I*) for a query at standard recall level *R*

$$I_R = \max_{R' \geq R} P(R')$$

To evaluate a method where *n* queries were tested, the *n* interpolated precision points at each standard recall point were averaged, showing the decline of precision as recall increases. It intractable to perform statistical hypothesis tests comparing methods represented by 11 different performance values so to facilitate statistical comparison between methods, the mean average precision (MAP) was also calculated across the queries (Equation 14). MAP is highly correlated with the area under a precision-recall curve and the single value metrics (rather than a curve with

11-points) enable simple hypothesis testing using a Student's paired *t* test (Smucker, Allan, & Carterette, 2007).

Equation 14: Average precision (AP) for a query where *N* is the total number of results, *d* is the number of relevant results, *P(k)* is the precision at k results and $\Delta r(k)$ is the change in recall from cut-off *k-1* and *k* (thus only permits precision at 'relevant' ranks to be averaged).

$$A \quad = \frac{1}{d} \sum_{k=1}^{N} P(k) \Delta r(k)$$

2.3 Results

2.3.1 Phenotype annotation

OMIM phenotype annotation statistics are shown in Table 5. Compared to the curated annotations, unquantified text mining assigned more phenotype terms to each disease on average but detected a far narrower range of different terms overall, which tended to be more general (closer to the root node). Possible reasons for this are that text-mining identifies more general terms where descriptive text is not sufficiently specific to distinguish between relevant sub-nodes of the ontology, and that additional general terms are identified in the text that annotation curators would consider redundant (due to applying a more specific term). These reasons would also explain the greater number of total annotations for text-mining, with the HPO terms sharing more OMIM phenotypes on average. Text mining also detected a lower proportion of the full range of HPO terms due to only using the OMIM text description as an input, whereas the curated annotations utilise additional data sources, such as published clinical studies and individual clinical experience. When the text-mined terms were quantified, it resulted in over double the total annotation count.

Table 5: Metrics for different methods of annotating the OMIM phenotype catalogue with HPO terms. The subscript *N* denotes the group of phenotypes captured by all annotation methods. The subscript *X* denotes those phenotypes exclusively captured by each annotation method (curated vs. text-mined). "Quantified" text-mined annotations reflect the frequency with which HPO terms are found in OMIM descriptions, whereas "unquantified" annotations have been reduced so the frequency of every term is 1.

Annotation method	**Phenotypes**	**Total annotations**	**Terms used**	**Average phenotypes per term**	**Average distance to root**
Curated	6,902	90,236	6,825	13.2	6.50
Text mining, unquantified	7,600	105,644	4,719	22.4	6.43
Text mining, quantified	7,600	230,274	4,719	22.4	6.43
Curated $_N$	6,518	88,533	6,765	13.1	6.50
Text-mined, unquantified $_N$	6,518	99,126	4,679	21.2	6.43
Text-mined, quantified $_N$	6,518	215,895	4,679	21.2	6.43
Curated $_X$	384	1,703	918	1.86	6.11
Text-mined, unquantified $_X$	1,082	6,518	1,598	4.08	6.19
Text-mined, quantified $_X$	1,082	14,379	1,598	4.08	6.19

2.3.2 Correlation between penetrance data and text-mined frequency

Where both term penetrance data and text-mined frequency data is available, there is a weak positive correlation between them (Figure 9). This becomes slightly stronger when removing the penetrance data described in the HPO by words such as "Hallmark", "Common", "Rare" which have approximate frequency definitions within the HPO documentation (P. N. Robinson & Mundlos, 2010).

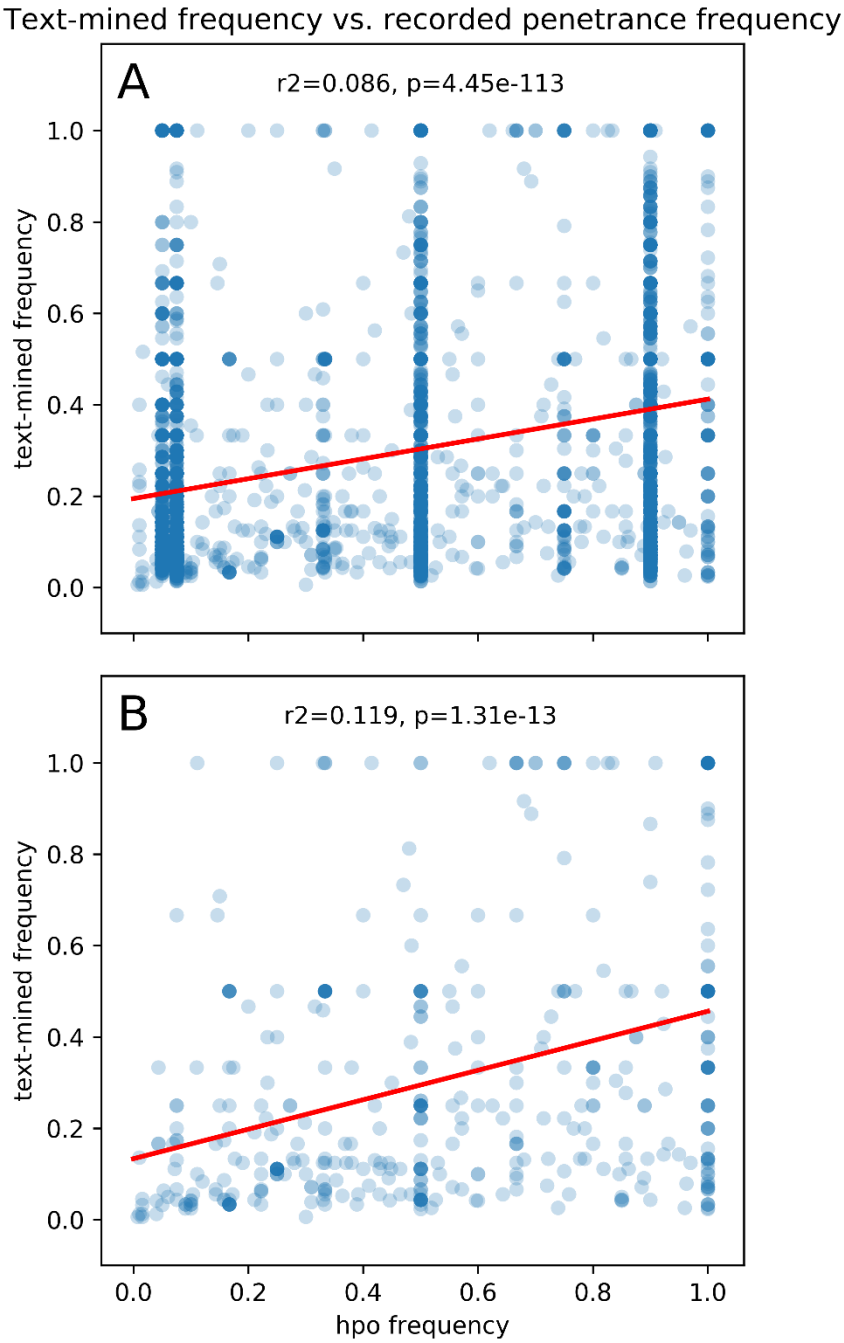


Figure 9: Correlation between HPO term penetrance statistics and derived text-mined frequency. Text-mined frequency was calculated by dividing term frequency by the highest HPO term frequency within the OMIM disease. A: Only congruent HPO annotations were considered (i.e. the HPO term was annotated to the OMIM entry in both the curated and text-mined dataset) and only OMIM entries that contained at least 5 quantified terms in both datasets were considered. This resulted in 5651 HPO annotations being tested, spanning 597 OMIM entries. B: HPO terms quantified in the curated dataset using text descriptions such as “Hallmark”, “Common”, “Rare”, etc. were excluded, resulting in 435 terms being tested, spanning 48 phenotypes.

2.3.3 Interpolated precision-recall and mean average precision (MAP) results

Two different similarity measures (cosine and $Resnik_{avg,max|sym}$) and three different phenotype annotation methods (curated, quantified text-mined and unquantified text-mined) were tested using the OMIM PS as a set of known similar phenotypes with which to simulate queries. Only diseases annotated by all methods (n=6,518; Table 5) were included in the analysis and diseases in multiple phenotypic series were not queried (n=47). 2,317 OMIM PS queries were used for analysis (Figure 10), which were a good representation of phenotypic spectrum of the overall OMIM disease catalogue (Figure 11). Queries were evaluated for each method using interpolated precision-recall (Figure 12; Equation 13) and mean average precision (Figure 13; Equation 14) was also used as a single value metric for each query. Mean average precision (MAP) was highly correlated with interpolated precision-recall (Figure 14) and enabled statistical hypothesis testing for comparison between methods using a paired Student's *t* test. *P*-values have been corrected for multiple testing under dependency (Benjamini & Yekutieli, 2001) unless stated otherwise.

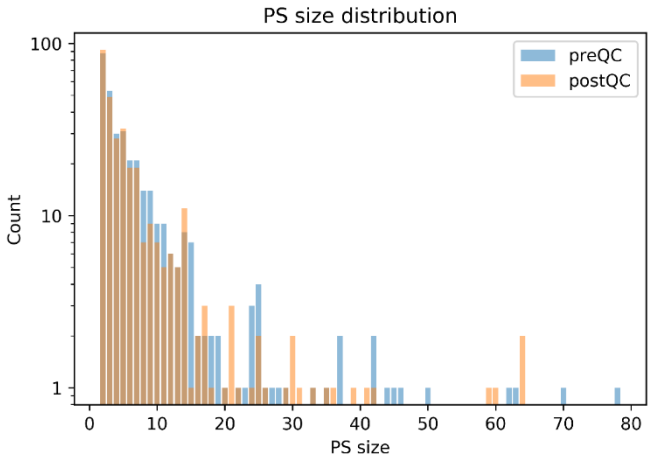


Figure 10: Distribution of number of OMIM phenotypes in each phenotypic series (PS) before (preQC; blue) and after (postQC; orange) removal of phenotypes in multiple phenotypic series. Method benchmarking was performed on the postQC OMIM PS group.

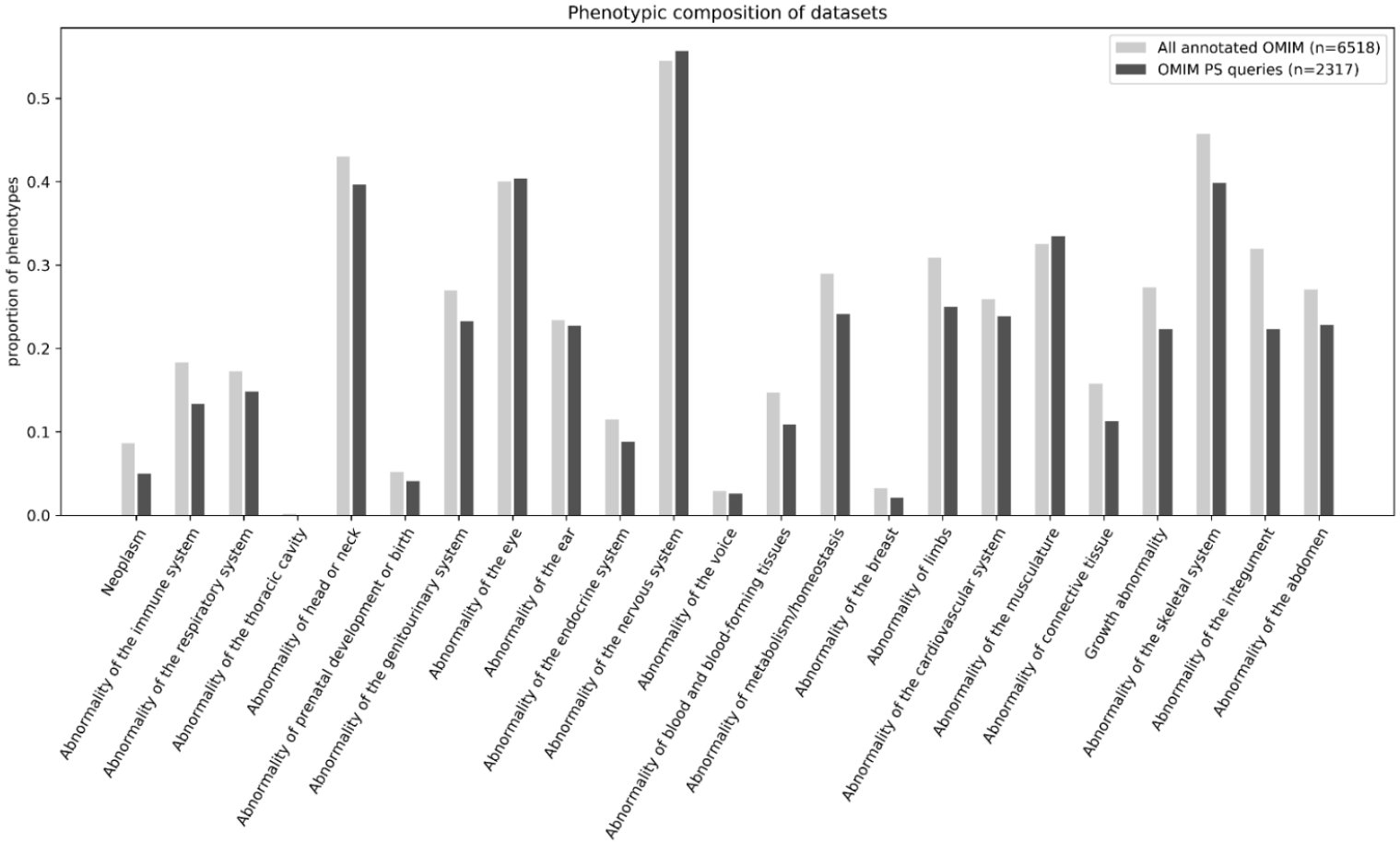


Figure 11: Range of disease phenotypes tested in the OMIM PS benchmarking, as compared to the full annotated OMIM catalogue (annotated by each method), characterised by the proportion of queries covered by each HPO term directly below 'Phenotypic abnormality' (HP:0000118).

When HPO annotations were not quantified (curated (c) and unquantified text mining (u)) the performance of cosine similarity had a modest but significant advantage over $Resnik_{avg,max|sym}$ ($P_c = 4.62 \times 10^{-13}$, $P_u = 4.47 \times 10^{-14}$), but cosine similarity was far superior when the HPO terms were quantified ($P_q = 2.83 \times 10^{-95}$) (Figure 12A; Figure 13A). When annotations were not quantified the similarity in performance between the two algorithms was expected due to their similar premise – $Resnik_{avg,max|sym}$ aggregates similarity in pairwise fashion while cosine is a groupwise method, but without quantification the only schema to weight terms is the specificity (IC).

When phenotype annotations are quantified, cosine similarity performs better due to its ability to down-weight the vector features of more general terms (which have a lower IC) and noise terms (which are likely to be found at a lower frequency to 'genuine' terms) that text mining is more prone to detecting. Using cosine similarity, the performance of each annotation method was assessed. Using these benchmarking tests and metrics, quantified text mining is superior to curated annotation ($P = 9.02 \times 10^{-58}$), although unquantified text mining is inferior to curated and quantified methods ($P = 2.38 \times 10^{-10}$ and $P = 1.02 \times 10^{-151}$, respectively). Having observed that querying a disease consisting of quantified phenotype terms against a quantified reference set was the optimal setting, the quantified reference set was tested against others, without quantification of the query terms.

The quantified reference disease set compares favourably to both the curated and unquantified text-mined reference sets when using unquantified text-mined queries ($P = 2.55 \times 10^{-64}$ and $P = 1.38 \times 10^{-63}$, respectively) (Figure 12B; Figure 13B). However, when querying with curated phenotype annotations, a quantified

reference set provides no clear benefit in comparison to the curated annotation set (P = 0.658[a], Figure 12C; Figure 13C). Curated queries were more effective with the curated reference set while text-mined queries were more effective with the text-mined reference sets.

[a] Without multiple testing correction

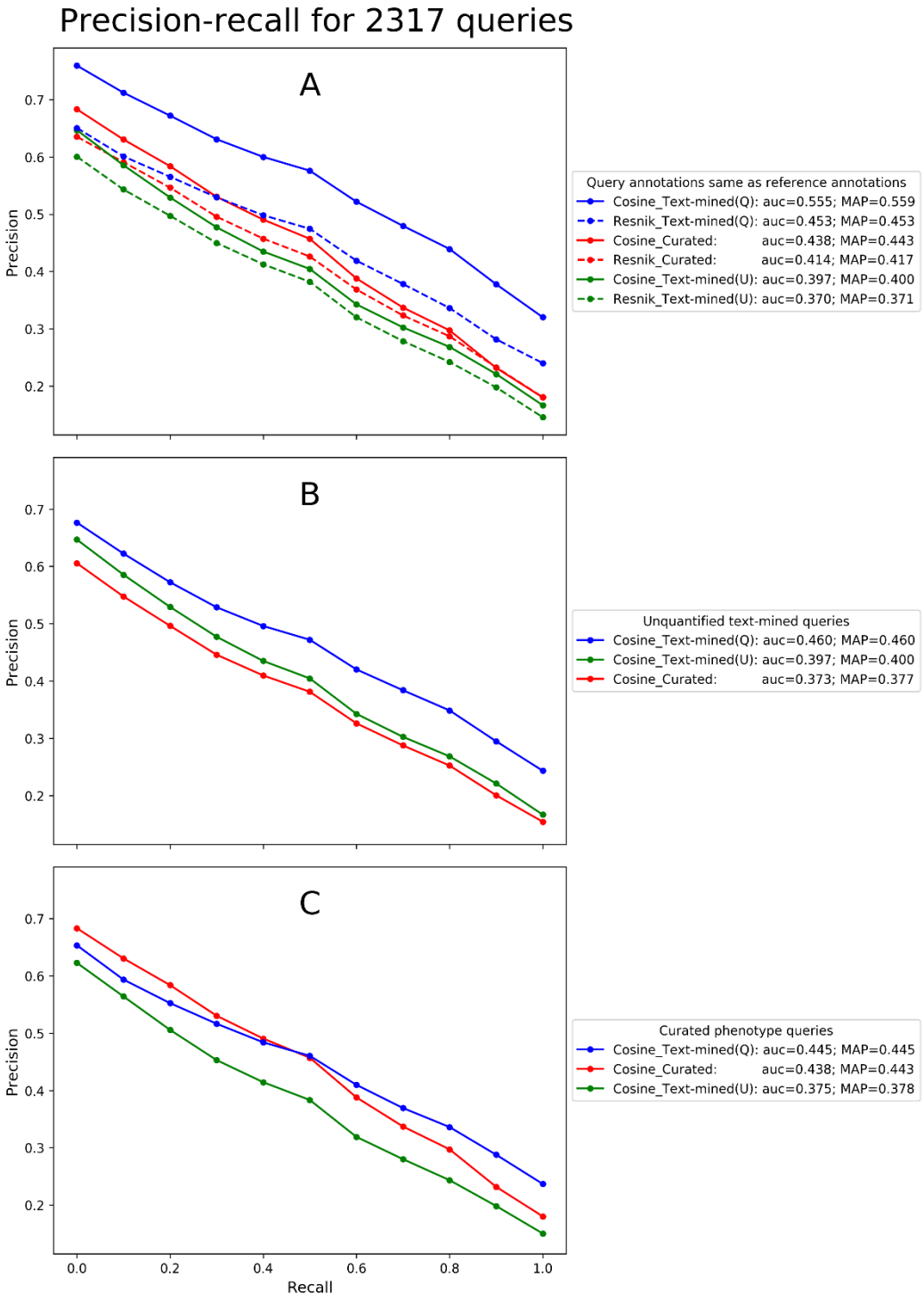


Figure 12: Average 11-point interpolated precision-recall for 2317 queries using different combinations of phenotype annotation and similarity methods. Similarity measure is denoted by linestyle (solid = cosine similarity; dashed = $\text{Resnik}_{\text{avg,max|sym}}$). Reference set annotation is denoted by line colour (red = HPO curated annotation; green = unquantified text mining; blue = quantified text mining). Area under the precision-recall curve and mean average precision (MAP) are indicated in the legend. A: All combinations of annotation and similarity measure were tested, keeping the annotation setting of the query the same as the annotation of the reference set. B: All reference annotation methods tested with queries from the unquantified text-mined set only. C: All annotation methods tested with queries from the curated set only. B&C: only the cosine similarity measure was displayed as it was superior to $\text{Resnik}_{\text{avg,max|sym}}$ in all cases.

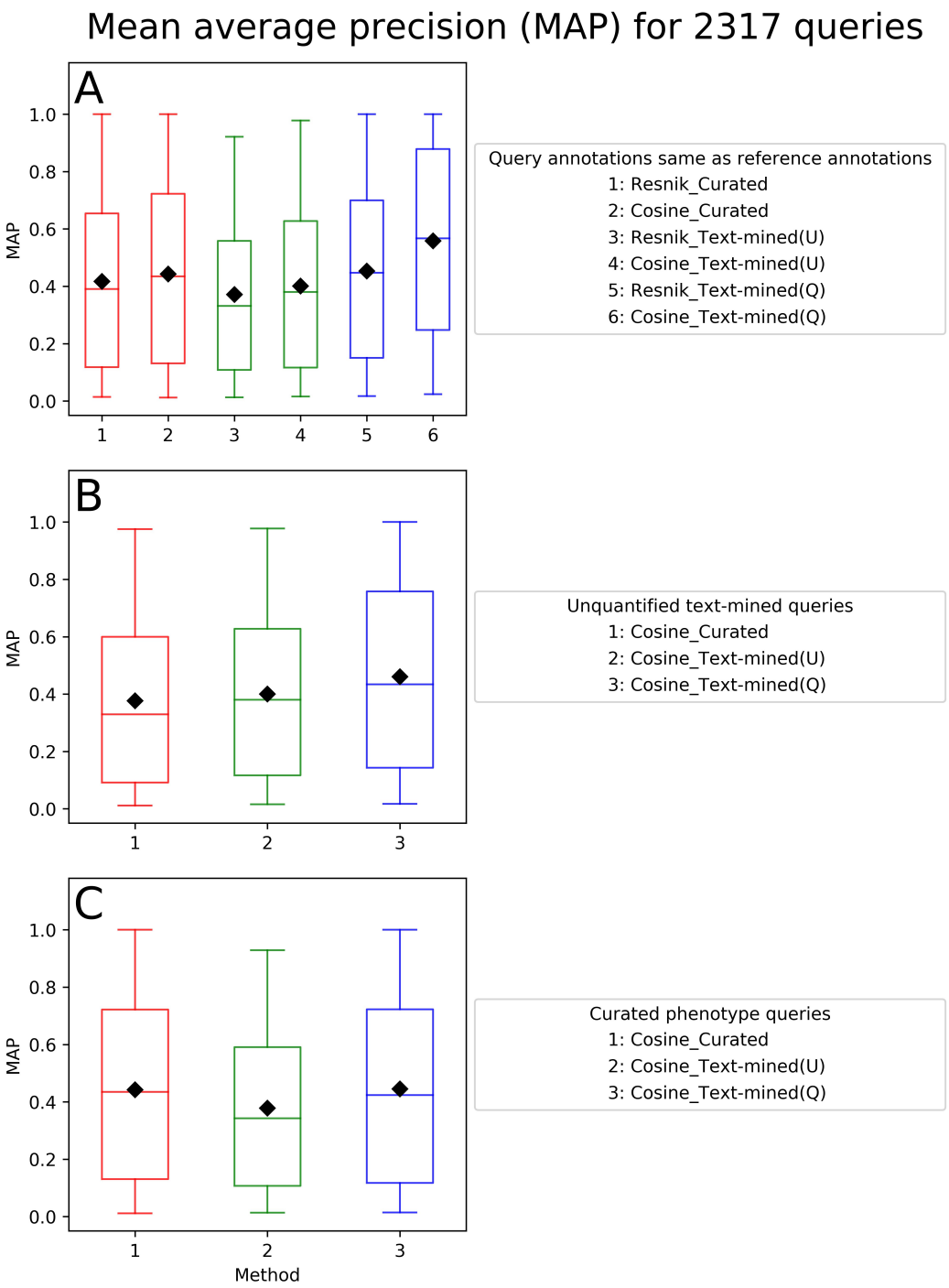


Figure 13: Mean average precision (MAP) plots for 2317 queries using different combinations of phenotype annotation and similarity methods (corresponding to the interpolated precision plots in Figure 12). A: All combinations of annotation and similarity measure were tested, keeping the annotation setting of the query the same as the annotation of the reference set. B: All reference annotation methods tested with queries from the unquantified text-mined set only. C: All annotation methods tested with queries from the curated set only. B&C: only the cosine similarity measure was displayed as it was superior to $Resnik_{avg,max|sym}$ in all cases. Box whiskers represent furthest data points within 1.5 × IQR (interquartile range).

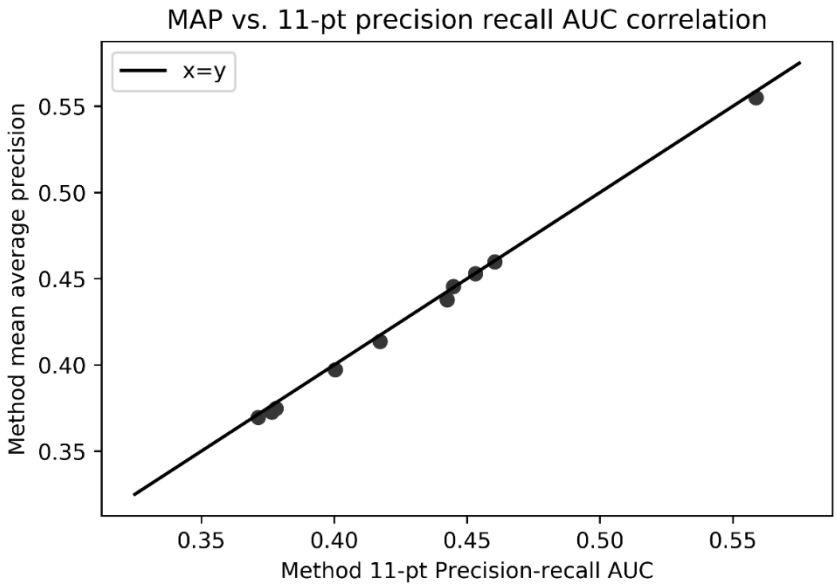


Figure 14: Correlation between MAP and 11-pt precision recall AUC for methods in Figure 12 and Figure 13.

2.4 Discussion

This chapter presents a method for the identification of similar diseases where the reference diseases were annotated by text mining for phenotype terms, which included term quantification to make optimal use of the most relevant phenotypic features. This approach offers a clear advantage over current methods that incorporate binary phenotype annotations that indicate only the presence or absence of clinical terms in genetic disease. Ideally, clinical terms could be quantified in the representation of a genetic disease using a full dataset on their prevalence among individuals having that disease. However, such a comprehensive dataset does not yet exist (41% of current HPO curated annotations have quantification information with 48% of diseases containing at least one quantified phenotype term) and it is unclear how some currently employed similarity methods can integrate knowledge of feature prevalence.

Whilst frequency extracted from text mining phenotypic descriptions is not expected to fully capture the penetrance of terms across affected cases, there is a weak positive correlation where term penetrance data is available (Figure 9) and it serves as an effective weighting scheme that encompasses all phenotypes. It is likely that the performance of this text-mining approach would further improve using additional relevant input phenotype descriptions, including for example, text input from OMIM literature references and further literature on MEDLINE. This method of annotation has obvious limitations – simplistic text mining is unable to account for the context in which disease terms are mentioned, and therefore will include noise due including terms mentioned following negation or when discussing/comparing to other phenotypes. Despite the limitations of this method

of annotation, this relatively simple method of term quantification aids the retrieval of similar diseases.

None of the tested methods incorporate HPO negative annotations, which have also been curated alongside the HPO disease phenotype annotations file. Although this is a potentially useful dataset, as it records phenotype characteristics that are NOT presented in genetic disease, it is unclear how similarity methods would incorporate this information. With the vector representation of diseases, phenotypic characteristics are already set to 0 if they aren't present in the annotation file – perhaps incorporating the negative annotations to set them to a negative number could improve performance. Negative annotations could also be used to augment the text-mined phenotype annotation file, helping to remove false positive annotations.

Additionally, a cosine similarity measure was investigated as an alternative to currently used phenotype similarity measures, as its construction is less sensitive to noise associated with quantitative annotation. $Resnik_{avg,max|sym}$ was selected as the measure to compare to due to its usage in both differential diagnostic and upstream variant prioritisation methods, although ideally these benchmarking methods should be additionally applied to BOQA and PhenoDigm methods of calculating phenotype similarity.

Validation involved testing the ability of different combinations of annotation and similarity methods to group known similar OMIM diseases closely to each other. Quantification of phenotypic information enhanced the ability to identify similar diseases, compared to both the curated annotation set and an unquantified version of the same reference set, when both the query and reference set are quantified.

Cosine similarity performed roughly as well as the $Resnik_{avg,max|sym}$ measure of similarity when unquantified reference sets were used, but greatly outperformed it when quantifying reference phenotype terms. It is less clear whether the use of a quantified reference set is beneficial when the query is not quantified, shown in the differential relative performance of the quantified reference set compared to the curated reference set when using either unquantified-text mined queries or curated queries. Curated queries performed better with the curated reference set and text-mined queries performed better with the text-mined reference sets. This represents a drawback in the validation of the methodologies, as the use of OMIM phenotypic series may not resolve potential biases within the reference phenotype annotation sets. There is a possibility that some curated annotations are copied across phenotypic series, and that text mining makes use of copied text within OMIM entries from the same phenotypic series. However, the use of OMIM phenotypic series as the known similar diseases enabled the methods to be tested across a considerable number of phenotypes (~35% of annotated OMIM records), representing a very diverse range of diseases. The validation of phenotypic similarity methods is always challenging, as until recently there has not been a large amount of real-world data to perform benchmarking on, so simulations or small patient datasets are used. In these datasets and simulations it has been standard to simulate noise (adding random terms) and imprecision (converting specific terms to more general terms) in the queries (Javed et al., 2014; Köhler et al., 2009; Smedley et al., 2014; Zemojtel et al., 2014). This has not been performed here – the high number (2,317) of queries for these benchmarking metrics, as well as that the OMIM phenotypic series contained diseases that were distinct [though

clinically similar] entities comprising slightly different phenotypic constellations, were considered sufficient for the simulation of noise.

Work in the next chapter involves the testing of these methods on real patient datasets, where consideration is given to the magnitude of the phenotype similarity score and its representation of the belief that a particular clinical diagnosis can be made and a particular disease gene identified. One drawback of the method presented here is that it does not currently incorporate a measure of statistical significance, as in the Phenomizer (Köhler et al., 2009), so it is unable to assess what level of cosine similarity represents a good match for the particular phenotype. However, to run accurate simulations of HPO queries in order to calculate P-values, the prevalence for each HPO term must be estimated rather than picking any HPO term randomly. The population prevalence (or rare disease population prevalence) of clinical phenotypes is difficult to estimate with data currently available – Orphanet has prevalence information on many rare disorders (Maiella et al., 2013) but these are constellations of phenotypic characteristics rather than the characteristics themselves.

Text mining for phenotype annotations can also have utility in a clinical context; its quick and systematic nature could make it highly valuable in large clinical genetics services. Manual assignment of clinical ontology terms has only recently become widely undertaken and is performed with variable degrees of diligence. As an alternative, the text mining of patient clinic letters for phenotypic terms would enable rapid and systematic definition of patient phenotypes. Term quantification would enable scoring of terms based on the belief that they are truly descriptive of the patient, and evidence presented here suggests that querying quantified phenotypes onto a quantified phenotype reference set improves performance of

phenotype similarity algorithms, so it may be worthwhile to pursue this. The incorporation of more sophisticated text mining features such as detection of term negation and modifiers for severity would also be of value. Text mining patient records that encompass a prolonged period would also enable longitudinal phenotype data to be collected. This is particularly pertinent in the context of syndromes where different clinical features appear at different ages. Patients could then be compared based on their clinical presentation at defined timepoints.

To summarise, this chapter demonstrates that quantifying clinical terms can be an effective method of refining phenotype descriptions, beyond a simple representation as binary observations. When calculating similarity, term frequency becomes an additional feature by which terms can be weighted rather than only their IC. A simple vector-based method has been utilised for calculating cosine similarity between quantified phenotypic definitions, which is able to consider term frequency and specificity, and this method shows improvement compared to currently employed methods in classifying related OMIM diseases as similar.

Chapter 3 - Using patient similarity to a reference set to predict disease genes in diagnosed cases

3.1 Introduction

The previous chapter was an investigation into a novel method of identifying similar disease phenotypes, through using different methods of phenotype annotation and quantifying similarity. In the following two chapters these methods are tested using the molecular diagnosis of real rare disease patients, to evaluate their ability to identify causative genes through identifying similar phenotypes to the patient.

This chapter utilises data released by a rare disease sequencing project from the DDD consortium, who undertook exome sequencing to attempt to identify the causative genetic variants of 1,133 undiagnosed patients with developmental disorders (The Deciphering Developmental Disorders Study, 2014; Wright et al., 2015). There have been several recent projects undertaking the sequencing of patients with rare disease for genetic diagnosis, either focussed on one phenotypic area (de Ligt et al., 2012; Y Yang et al., 2014; Yaping Yang et al., 2013) or across multiple broad phenotypic areas (H. Lee et al., 2014; Sawyer et al., 2016; Taylor et al., 2015; Wright et al., 2015). Standard variant filters (frequency, consequence, zygosity, predicted pathogenicity) are not sufficient in achieving a diagnosis (Bamshad et al., 2011), so genetic variants are either further filtered using a primary gene panel for the phenotypic area (H. Lee et al., 2014; Wright et al., 2015) or all remaining variants are investigated for causality (de Ligt et al., 2012; Sawyer et al., 2016; Taylor et al., 2015; Y Yang et al., 2014; Yaping Yang et al., 2013).

Diagnostic yields were consistently reported to be between 25 and 30% (Schwarze et al., 2018), leaving the majority of patients without a diagnosis. However, it is difficult to assess and compare diagnostic rates of rare disease sequencing projects, as they may cover different phenotypic areas (where a diagnosis may be easier/harder to achieve) and some datasets may contain cases which have been previously investigated without a positive finding (and therefore enriched for 'difficult' cases, that indeed may not be monogenic) (The Deciphering Developmental Disorders Study, 2014).

All 1,133 DDD patients were children (median age 5.5 years) recruited to the DDD study by their UK NHS or Irish Regional Genetics Service, where patient HPO terms were also recorded. The most commonly presented phenotypic features were intellectual disability or developmental delay (87% of children), abnormalities revealed by cranial MRI (30%), seizures (24%) and congenital heart defects (11%). The DDD diagnosis strategy involved exome sequencing all patients (as well as performing aCGH on 1,009) to identify SNVs, indels and CNVs. Standard variant filters were used to exclude common (>1% minor allele frequency) and non-functional (not protein-altering) variants. Variants were then filtered by gene, retaining only those within genes in the Developmental Disorders Genotype-to-Phenotype (DDG2P) database (Firth et al., 2009), a continuously updated list spanning over 1,000 genes causative of developmental disorders – the November 2013 version used includes 1,128 genes. Each remaining variant was evaluated with respect to its potential causality of the patient phenotype, considering whether the patient phenotypic presentation is consistent with phenotypes/syndromes caused by the gene, as well as the genetic mechanisms previously reported for diseases caused by the gene (autosomal dominant, autosomal recessive, X-linked)

and whether this is consistent with the variant consequence on gene product (loss of function, activating mutation, increased gene dosage, etc.) (Wright et al., 2015). 317 patients were reported a likely diagnosis of a SNV, indel or CNV, and a further 35 patients were reported to have a pathogenic mutation in a novel gene linked to developmental disorders following functional validation.

The phenotype querying methods tested in the previous chapter were tested again here on the patients, evaluating their ability to prioritise the correct diagnostic gene based on the phenotypic presentation of the patient (using the recorded HPO terms). Analysis was initially based on ranks of causative genes, enabling comparison between methodologies to be undertaken. Further work was carried out to assess the level of belief assigned to the causative gene, based on the magnitude of the similarity score rather than the rank. This enabled methodology to be compared across different phenotype query methods, but also against the gene panel method employed – the gene panel was considered a uniform distribution of belief across the ~1,000 genes in the panel, whereas the results of phenotype similarity searching confer different levels of belief for each gene. A phenotype query method would be considered superior to a gene panel approach if the assigned probability of the causative gene was greater than drawing it randomly from all genes in the panel. This work involved the rescaling of phenotype similarity scores to value reflecting the probability of phenotype similarity, rather than assuming a linear relationship – a logistic function was selected based on simulated data from the previous chapter.

3.2 Materials and Methods

3.2.1 DDD data

The patient genetic diagnosis data released by the DDD contained 411 reported variants, belonging to 351 patients – this included 17 digenic diagnoses where two different genes containing pathogenic variants were reported, and the patient presented a composite phenotype. The remaining excess of reported variants were due to compound heterozygosity, hence two variants reported rather than a single homozygous variant. The 317 diagnoses and 35 novel genes added to 351 patients with reports (rather than 352) because one patient possessed both a diagnostic variant and a variant in a novel gene.

For straightforward benchmarking, these 351 patient diagnostic reports were filtered to 283 patients with pathogenic variants in a single gene (rather than digenic diagnoses, CNVs, UPDs or mosaicisms). These monogenic patients were further filtered to 258, whose gene mapped to an OMIM phenotype contained in the intersection of phenotypes covered by all 3 reference sets tested (Table 5).

Table 5 (repeated): The three OMIM reference disease HPO annotation sets tested in this chapter. Only OMIM diseases captured by all annotation methods, denoted by subscript *N*, were queried against. The subscript *X* denotes those phenotypes exclusively captured by each annotation method (curated vs. text-mined). "Quantified" text-mined annotations reflect the frequency with which HPO terms are found in OMIM descriptions, whereas "unquantified" annotations have been reduced so the frequency of every term is 1.

Annotation method	Phenotypes	Total annotations	Terms used	Average phenotypes per term	Average distance to root
Curated	6,902	90,236	6,825	13.2	6.50
Text mining, unquantified	7,600	105,644	4,719	22.4	6.43
Text mining, quantified	7,600	230,274	4,719	22.4	6.43
Curated $_N$	6,518	88,533	6,765	13.1	6.50
Text-mined, unquantified $_N$	6,518	99,126	4,679	21.2	6.43
Text-mined, quantified $_N$	6,518	215,895	4,679	21.2	6.43
Curated $_X$	384	1,703	918	1.86	6.11
Text-mined, unquantified $_X$	1,082	6,518	1,598	4.08	6.19
Text-mined, quantified $_X$	1,082	14,379	1,598	4.08	6.19

3.2.2 Reference disease annotations and similarity methods

For these 258 patients, their HPO terms were queried to the same three reference disease annotation sets from the previous chapter (curated, quantified text-mined and unquantified text-mined; Table 5), again only querying against the intersection of phenotypes covered by them (n=6518).

As with the benchmarking in the previous chapter, the performance of $Resnik_{avg,max|sym}$. Additionally, BOQA (Bauer et al., 2012) was tested using all reference sets (all combinations of similarity and annotation methods listed in Table 6). BOQA is able to incorporate frequency information, so the frequencies associated with curated HPO annotation frequencies were also used. To use BOQA with a compatible representation of text-mined phenotype annotation frequencies,

term counts were converted to derived penetrance statistics by dividing the count of each annotated HPO term by the highest HPO count within the phenotype.

Equation 3 (repeat): Resnik similarity between two terms (Resnik, 1999).

$$sim(t1, t2) = \quad (MICA(t1, t2))$$

Equation 6 (repeat): Best-match average term similarity between diseases Q and D

$$sim(Q \rightarrow D) = avg\left[\sum_{t1\ \epsilon\ Q} \max_{t2\ \epsilon\ D} sim(t1, t2)\right]$$

Equation 7 (repeat): Symmetrical similarity between diseases Q and D.

$$sim_{symmetric}(Q, D) = \frac{1}{2} sim(Q \rightarrow D) + \frac{1}{2} sim(D \rightarrow Q)$$

Equation 9 (repeat): Cosine similarity between vectors Q and D

$$sim(Q, D) = \cos\theta = \frac{Q \cdot D}{\|Q\| \cdot \|D\|}$$

Table 6: Methods tested in DDD benchmarking.

Annotation method	Similarity method
Curated Text-mined (unquantified) Text-mined (quantified)	$Resnik_{avg,max\|sym}$
Curated Text-mined (unquantified) Text-mined (quantified)	Cosine
Curated Curated (quantified) Text-mined (unquantified) Text-mined (quantified)	BOQA

3.2.3 DDG2P genes

The DDG2P gene panel (Firth et al., 2009) used in the initial DDD consortium diagnostic workflow was the November 2013 version, containing 1,128 genes (Wright et al., 2015). For this benchmarking, an updated version of the DDG2P panel was used from January 2016 (downloaded: 05/01/16), containing 1,313 genes (encompassing 1,814 gene-phenotype relationships), which was presumed to contain all the novel gene-phenotype relationships presented in the initial publication (The Deciphering Developmental Disorders Study, 2014). This panel was filtered to 1,268 genes that possessed causal relationship mappings to OMIM disorders, and could therefore be ranked/scored in relation to the patient phenotypes.

3.2.4 Mapping phenotype scores to DDG2P genes – rank analysis

After querying a patient phenotype with a particular query method, 6,518 phenotype scores are returned. These similarity scores were converted to gene scores for each gene in the DDG2P list (n=1,268) by taking the *highest* score mapping to each gene (some OMIM phenotypes map to multiple genes). The rank of the causative gene for each patient is then able to be compared across multiple methods.

3.2.5 Logistic function development – score analysis

Further analysis considered the magnitude of similarity score returned by similarity methods for each gene within the DDG2P panel, and how predictive these were of the causative gene beyond selecting the gene from the panel at random. Rather than assuming a linear relationship between similarity score and probability of similarity, data from the first chapter was used to establish the relationship between similarity score and probability of similarity. For each annotation and similarity method, each phenotype within an OMIM phenotypic series (PS) was queried back to the reference set, using similarity scores between phenotypes in the same series as examples of true positive similarity scores, and scores between all other phenotypes as examples of false positives. At evenly spaced bins (n=100) throughout the range of similarity scores, the proportion of true positive matches, *T*, was calculated (Equation 15). These were plot (Figure 15), and a generalised logistic function was selected to fit the data (Equation 16).

Equation 15: Calculation of *T* for each bin – TP is the number of scores between "true positive" matches within the bin; ALL is the total number of scores within the bin.

$$T_{bin} = \frac{T_{bin}}{ALL_{bin}}$$

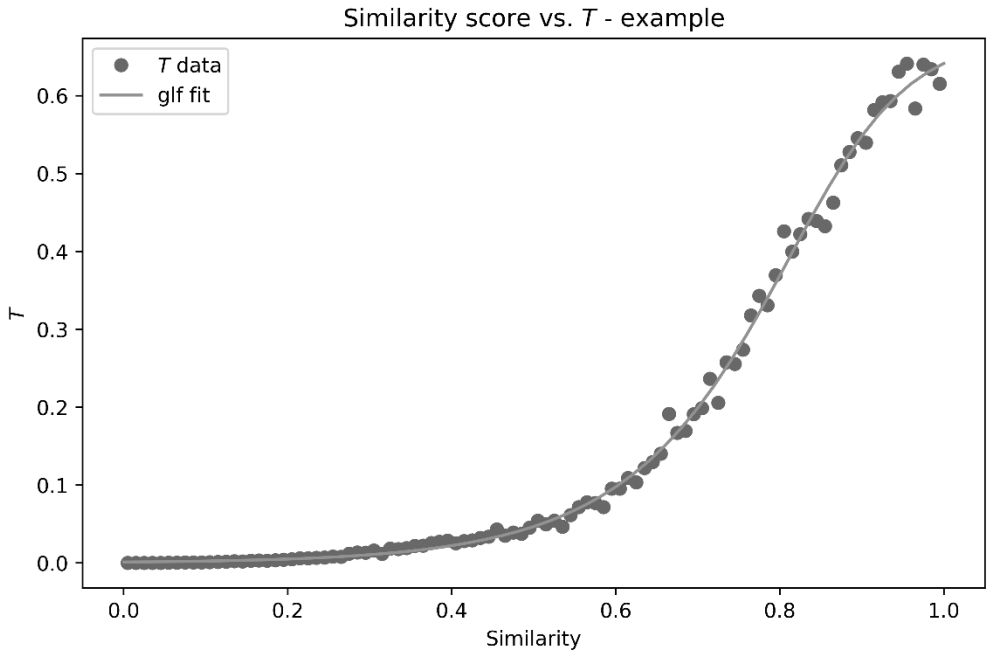


Figure 15: *T*, the proportion of true positives for each similarity score bin, with the generalised logistic function fit. Quantified text-mined phenotypes were queried in this example.

Equation 16: Generalised logistic function fit to the *T* data using midpoint *x* of each bin. Data was fit to using non-linear least squares, optimising variables *K*, *Q*, *B*, *M* and *v*. *K*: the upper asymptote; *Q*: fixes the point of inflection; *B*: the growth rate; *M*: the point of maximum growth; *v*: asymmetry parameter.

$$T = \frac{K}{(1 + Qe^{-B(x-M)})^{1/v}}$$

The generalised logistic function was fit for each method (apart from BOQA which is already calculated probabilistically) and used to rescale the phenotypic similarity scores to reflect probability of similarity. These rescaled similarity scores were then converted to DDG2P gene scores using the highest disease score that mapped to each gene, and normalised so all gene scores summed to one for each patient.

3.3 Results

3.3.1 Logistic function optimisation

For each combination of annotation and similarity methods, a logistic function was fit to the fraction of true positive scores within each similarity bin, using phenotypes within OMIM PS as true positive matches (Figure 16, Table 7).

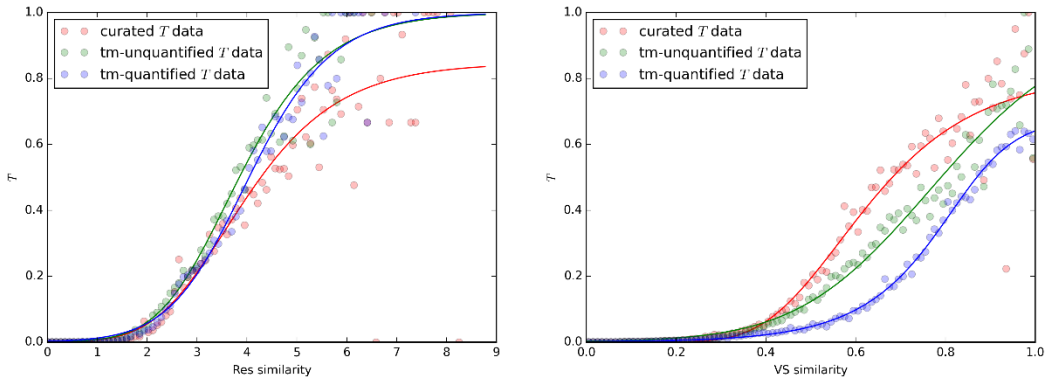


Figure 16: Logistic function fit to the fraction of true positive scores within each similarity bin (T) for each combination of annotation and similarity methods, using phenotypes within OMIM PS as true positive matches. Left: $Resnik_{avg,max|sym}$; Right: cosine similarity; Red: curated; Green: text-mined (unquantified); Blue: text-mined (quantified).

Table 7: Optimised generalised logistic function variables from Equation 16 for each combination of annotation and similarity method. Functions are plot in Figure 16.

Annotation method	**Similarity method**	***K***	***Q***	***B***	***M***	***v***
Curated		0.847	6.92	0.832	-0.626	0.21
Text mining, unquantified	Resnik	1	3.5	0.965	1.15	0.334
Text mining, quantified		1	1.54	1.09	3.1	0.63
Curated		0.81	0.515	6.15	0.207	0.0569
Text mining, unquantified	Cosine	1	0.628	5.46	0.753	0.598
Text mining, quantified		0.678	2.17	15.1	0.807	2.02

3.3.2 DDD patients

Of these 351 patient diagnostic reports, 283 were due to pathogenic variants in a single gene (rather than digenic diagnoses, CNVs, UPDs or mosaicisms). 258 of these monogenic diagnoses were in a gene that mapped to an OMIM phenotype which was contained in the intersection of phenotypes covered by all 3 reference

sets tested (Table 5). This dataset covers developmental disorders – 49% of patients had global developmental delay, as well as roughly 20% presenting microcephaly, delayed speech and language development, muscular hypotonia or intellectual disability. 81% of these patients were annotated with at least one of these top 5 terms (Table 8). Compared to the OMIM PS benchmarking, the DDD tests a narrower phenotypic spectrum of patients, heavily skewed towards 'Abnormality of the nervous system', 'Abnormality of head or neck' and 'Abnormality of the skeletal system' (Figure 17).

Table 8: Most common 15 HPO terms found within the monogenic DDD patient dataset (n=258). Terms were limited to only those within two nodes from the end of an ontology branch so only the most specific terms are displayed.

HPO term	*Description*	*Patient count (/258)*
HP:0001263	Global developmental delay	127
HP:0000252	Microcephaly	61
HP:0000750	Delayed speech and language development	54
HP:0001252	Muscular hypotonia	47
HP:0001249	Intellectual disability	43
HP:0100886	Abnormality of globe location	33
HP:0200006	Slanting of the palpebral fissure	32
HP:0001999	Abnormal facial shape	29
HP:0001328	Specific learning disability	27
HP:0000494	Downslanted palpebral fissures	23
HP:0000729	Autistic behavior	22
HP:0000717	Autism	21
HP:0001290	Generalized hypotonia	21
HP:0000486	Strabismus	20
HP:0000539	Abnormality of refraction	20

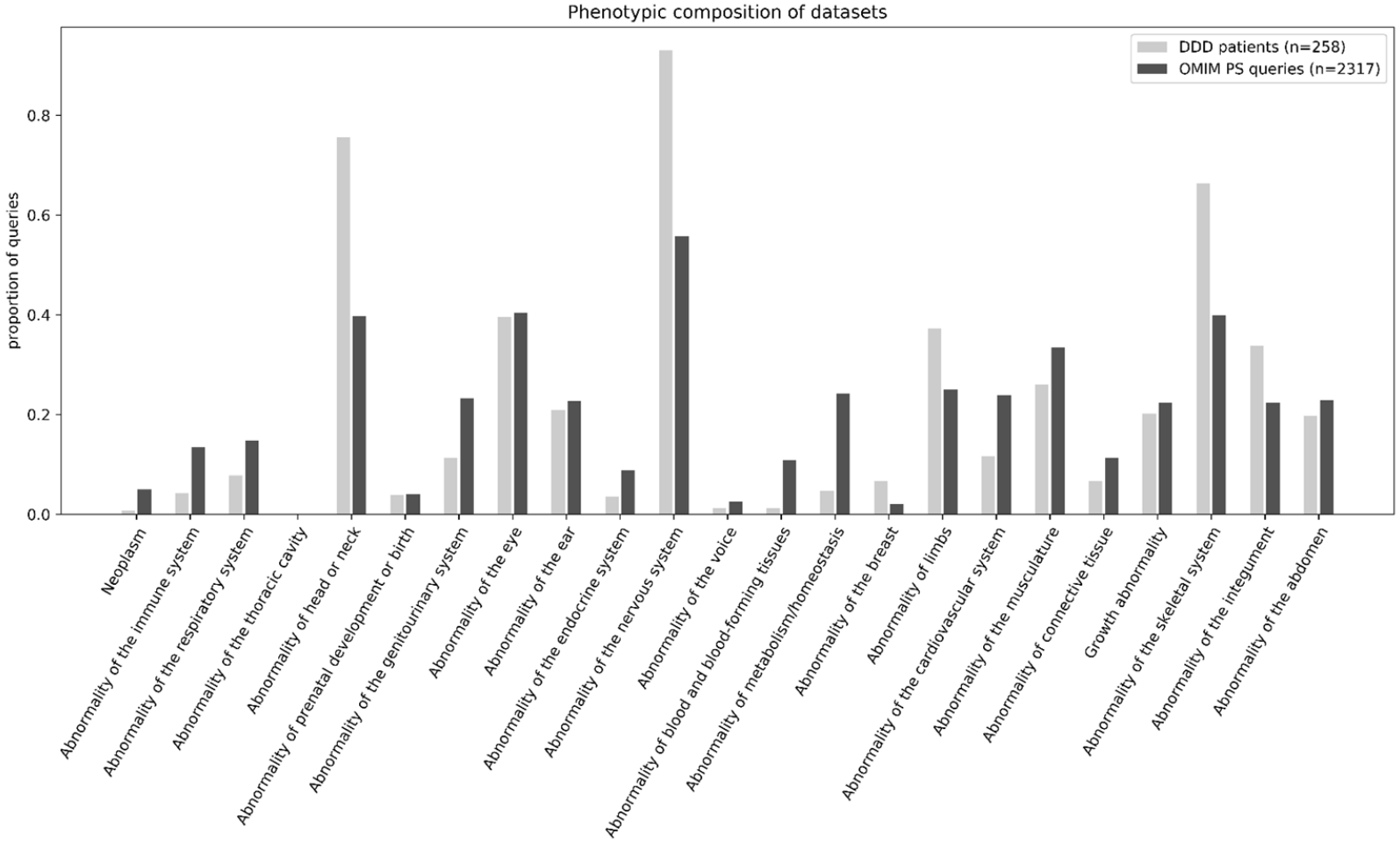


Figure 17: Range of disease phenotypes tested by the two different benchmarking strategies (DDD patients and OMIM PS) , characterised by the proportion of queries covered by each HPO term directly below ‘Phenotypic abnormality’ (HP:0000118).

3.3.3 Correct gene rank analysis

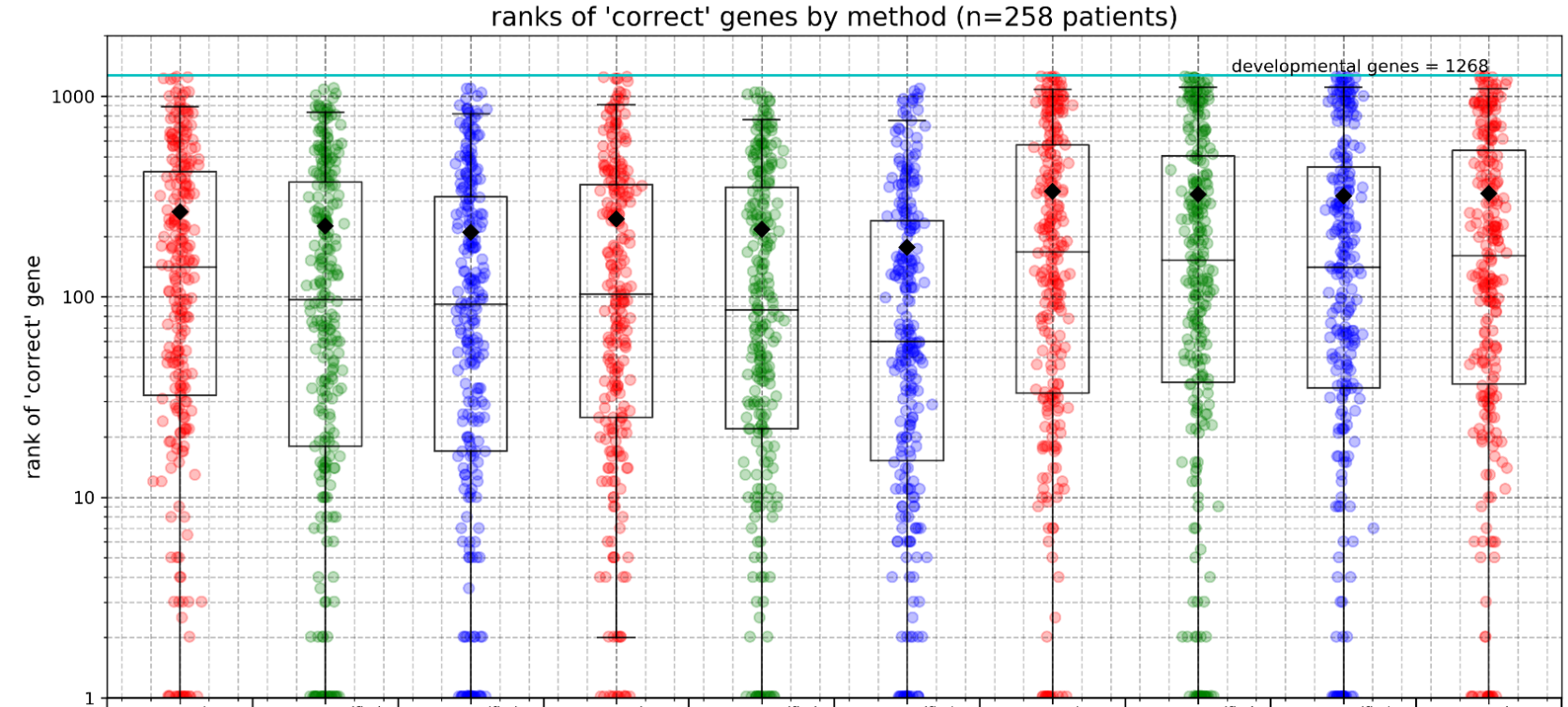
The different combinations of reference disease sets and phenotype similarity calculation methods (Table 6) were used to rank genes in the DDG2P list (used as a virtual panel in the original DDD study (Wright et al., 2015)) to determine which methods were most effective in prioritising genes within this large gene panel. After querying the 258 patients' HPO terms to the different OMIM disease reference sets, the disease similarity scores were converted to DDG2P gene scores using the OMIM gene-disease mappings (taking the top disease score for genes that cause multiple OMIM diseases). The ranks of the correct gene for each of the 258 patients were compared across methods using the Wilcoxon test followed by adjustment for multiple testing under dependence.

There was no significant difference between $Resnik_{avg,max|sym}$ and cosine similarity with the unquantified text-mined reference set) (Figure 18) ($P = 0.481$[a]), and with the curated reference set the improvement was borderline significant after multiple testing correction ($P = 0.0210$) – this is contrary to the finding in the OMIM PS benchmarking in the previous chapter where improvements conferred by vector-based cosine similarity were significant, though they were only marginal improvements. However, using cosine similarity resulted in an improvement for the quantified text-mined reference set ($P = 1.57 \times 10^{-4}$), which was consistent with the OMIM PS findings of the previous chapter. Using cosine similarity, comparison between reference sets showed no significant difference between the unquantified text-mined annotation set ahead of the curated annotations ($P = 0.36$[a]), but the quantified set showed a significant improvement in correct phenotype ranks

[a] Without multiple testing correction

over the curated set ($P = 1.39 \times 10^{-4}$). The quantified text-mined reference set also showed a significant improvement in comparison to its unquantified counterpart ($P = 1.03 \times 10^{-7}$), again supporting observations in the previous chapter. Using quantified text mining in combination with cosine similarity showed dramatic improvements over using BOQA with both the same quantified text-mined reference set, as well as the curated set for which BOQA could utilise penetrance statistics ($P = 1.85 \times 10^{-6}$ and $P = 8.62 \times 10^{-6}$ respectively). There was no significant difference between using BOQA with and without the incorporation of phenotype frequency data for both the curated and text-mined reference sets ($P = 0.325$[a] and $P = 0.13$ respectively). Interestingly, for each reference set, using both BOQA and $\text{Resnik}_{\text{avg,max|sym}}$ similarity methods predicted more correct genes at rank 1 than using vector space, although this trend is reversed at rank 10, where vector space becomes more sensitive than other similarity methods.

[a] Without multiple testing correction



	curated Res	tm-unquantified Res	tm-quantified Res	curated VS	tm-unquantified VS	tm-quantified VS	curated boqa	tm-unquantified boqa	tm-quantified boqa	curated-q boqa
median	141.0	97.0	92.0	103.0	86.0	60.0	167.2	151.8	140.0	160.5
mean	265.6	225.1	209.1	244.2	216.6	176.6	335.9	323.9	317.4	326.7
solved@1	16	25	25	12	19	18	15	22	23	15
solved@10	32	48	52	35	48	59	31	39	39	30
solved@100	115	131	135	128	135	145	99	104	115	101

Figure 18: Ranks of the ‘correct’ gene for 258 queries from the diagnosed DDD patient dataset, for the different combinations of reference annotations and query methods. Only phenotypes in all reference sets were queried (n=6518) and phenotype ranks were converted to DDG2P gene ranks (n=1268). Boxplot limits represent the 5th and 95th percentiles; black diamond indicates the mean rank.

3.3.4 Correct gene score-based analysis

Further analysis attempted to rescale the phenotype similarity scores to gene scores that reflected the probability of causality for each of the genes in the DDG2P list. After querying the patient HPO terms using the different methods, the generalised logistic function (Equation 16 with optimised variables for each method delineated in Table 7) was used to rescale the phenotypic similarity scores. BOQA outputs a probability for each disease and therefore rescaling wasn't used. The rescaled similarity scores were then converted to DDG2P gene scores using the OMIM gene-disease mappings (taking the top phenotype score for genes that cause multiple OMIM diseases) and normalised to sum to one to give an estimate of the probability for each gene. The gene scores of the causative gene for each patient were plotted against a baseline of the probability of selecting a random gene from the DDG2P list (Figure 19). The rescaled correct gene scores for each of the 258 patients were compared across methods using the Student's t test followed by adjustment for multiple testing under dependence.

For each reference set, cosine similarity outperformed Resnikavg,max|sym similarity in assigning a higher average probability to the disease-causing gene (Figure 19, Table 9), although this advantage is significant only for the curated ($P = 2.03 \times 10\text{-}3$) and the quantified text-mined reference sets ($P = 1.91 \times 10\text{-}3$). Compared to vector space and Resnikavg,max|sym, using BOQA to measure patient similarity to the respective reference sets resulted in a much higher mean probability assigned to the correct gene, which was due to a handful of outlying patients having a very high probability assigned to the causative gene (Each BOQA method had 15-25 patients with $\Delta>0.2$ for the correct gene). However, both BOQA and Resnikavg,max|sym approaches resulted in a low median Δ and performed

poorly for the majority of patients (in the best case, 61 of 258 patients had a positive Δ). This was also the case when querying the curated reference set using cosine similarity, which achieved a high average probability but was also offset by poor median performance, with over half of the correct genes having a lower probability than selecting the gene at random from the DDG2P list. The quantified reference set combined with vector-based cosine similarity space achieved the highest median probability for the correct genes. The quantified reference sets also achieved the highest number of patients with a positive Δ (probability subtracted by prior) for the correct gene. Additionally, the Resnikavg,max|sym similarity measure has a slight advantage over vector space in this respect. Again, the quantification of both curated and text-mined reference phenotypes made no significant improvement to BOQA ($P = 0.664$ and $P = 0.911$ respectively).

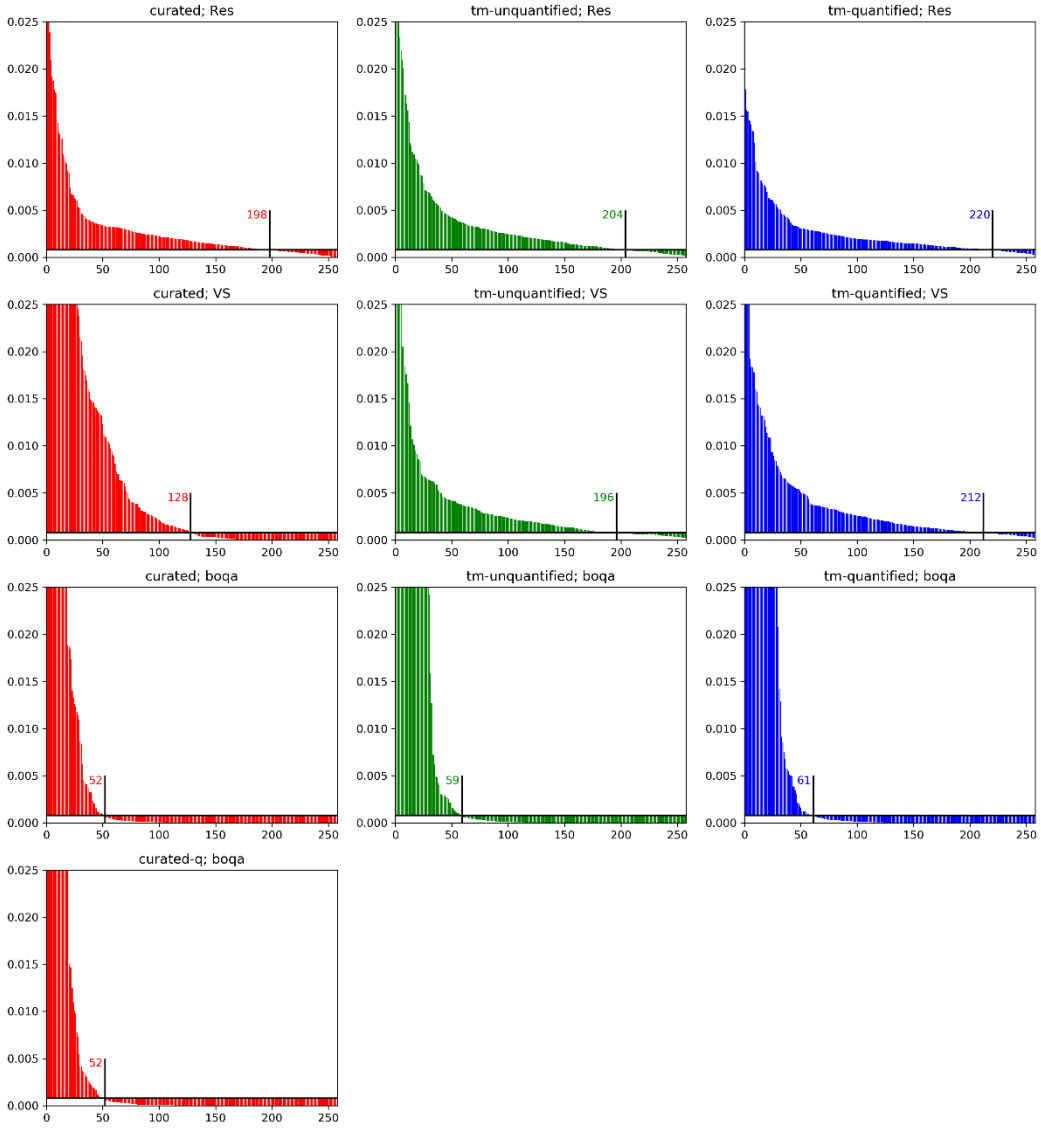


Figure 19: Probability (after logistic function rescaling) assigned to the correct gene for 258 DDD patient queries to different reference sets using different query methods. Probability was plot against a baseline of selecting a DDG2P gene at random (1/1268). Number of patients for which probability is higher than randomly selecting a DDG2P gene is indicated on the plot.

Table 9: Metrics for score-based analysis (after logistic function rescaling) of different phenotype annotation and similarity methods used to identify the causative genes for diagnosed DDD patients (from Figure 19). Δ = probability – prior; Fold change = Δ/prior.

Annotation method	Similarity method	Average Δ	Average fold change	Median Δ	Median fold change	n(Δ > 0)
Curated	Resnik	2.31E-03	2.927	9.26E-04	1.175	198
Text mining, unquantified		2.69E-03	3.414	1.12E-03	1.422	204
Text mining, quantified		1.88E-03	2.388	9.01E-04	1.143	220
Curated	Vector space	2.27E-02	28.731	-5.86E-05	-0.074	128
Text mining, unquantified		2.79E-03	3.532	1.05E-03	1.327	196
Text mining, quantified		3.43E-03	4.345	1.17E-03	1.479	212
Curated	BOQA	5.33E-02	67.529	-7.79E-04	-0.987	52
Text mining, unquantified		8.31E-02	105.426	-7.69E-04	-0.975	59
Text mining, quantified		7.65E-02	97.020	-7.54E-04	-0.956	61
Curated, quantified		5.21E-02	66.053	-7.69E-04	-0.975	52

3.3.5 Correlation between different benchmarking metrics across methods

Method performance in this benchmarking procedure was assessed using both ranks and scores of causative genes, and within the score-based analysis both median and average scores were considered. There was a negative correlation between the median and average causative gene scores for each method (Figure 20) – some methods were able to identify the correct causative gene with very high confidence in a small subset of patients but performed poorly on the rest, while other methods were able to prioritise the majority of causative genes, but with less confidence. The latter group of methods were also stronger at ranking genes, with a negative correlation between median gene score and gene rank, whilst the opposite is true for the former (Figure 21).

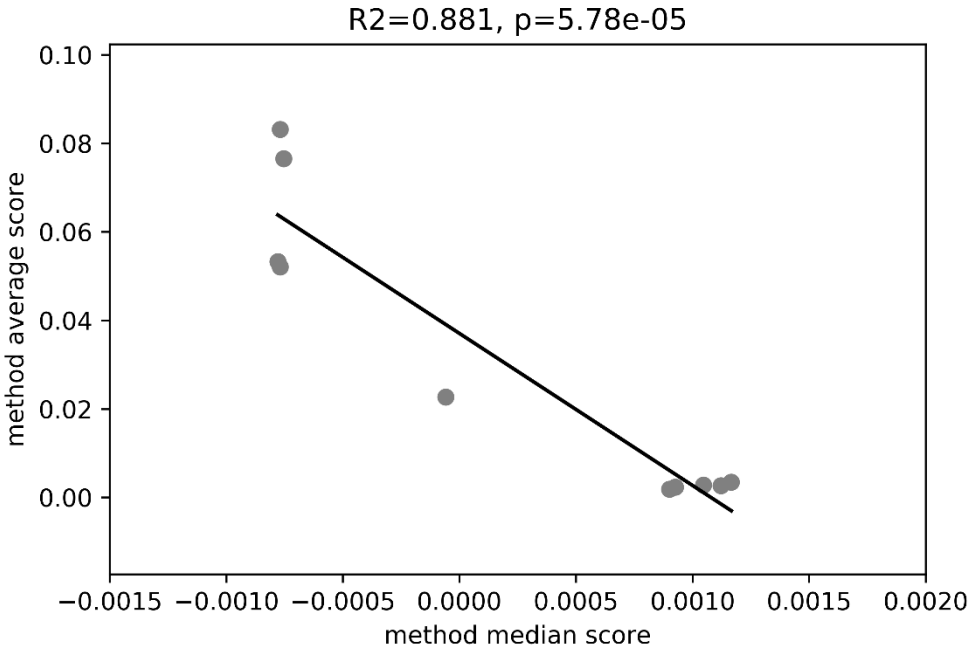


Figure 20: Negative correlation between method average causative gene score and median gene score from the DDD patient benchmarking (all methods in Table 6).

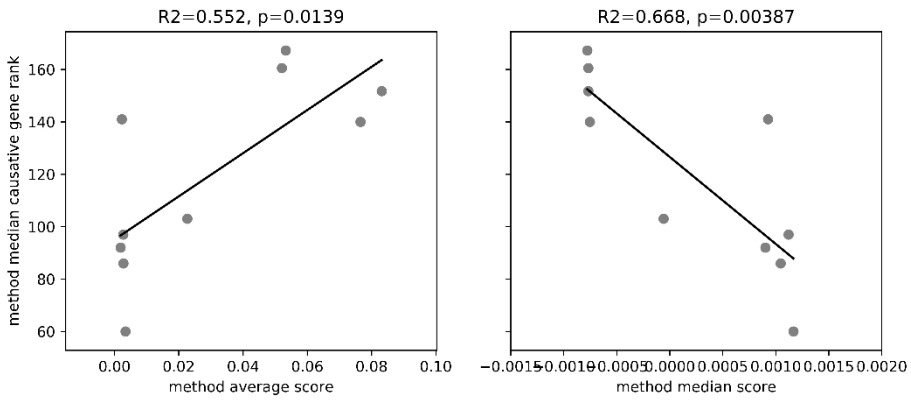


Figure 21: Correlations between different method evaluation metrics. Left – Negative correlation between method performance as measured by the median causative gene rank and average causative gene score (for all methods in Table 6); Right – Positive correlation between method performance as measured by the median causative gene rank and method average score .

3.4 Discussion

In this chapter the phenotype similarity methods tested in the previous chapter were applied to real patient dataset – those of the DDD consortium project (The Deciphering Developmental Disorders Study, 2014). Results were largely concordant between these tests and the OMIM PS benchmarking of the previous chapter, specifically in the correct gene rank analysis (Figure 18) and the median correct gene score (Table 9) – the use of cosine similarity was superior to $Resnik_{avg,max|sym}$ when reference disease phenotype annotations were quantified, and that the text-mined disease reference set was superior to both the unquantified text-mined and the curated reference set at prioritising candidate genes. These improvements were apparent despite the query HPO terms not having any quantification. Whilst the OMIM PS testing may have contained bias, as demonstrated by the relatively strong performance when the reference set and query were both either curated or text-mined, the DDD dataset contains no such bias as it consisted of real patients with HPO terms assigned *before* diagnosis. It is impossible to estimate what bias may be associated with the assignment of HPO terms to patients – clinicians may phenotype patients with a particular diagnosis under consideration, possibly leading to an ascertainment of phenotypic annotations from that particular disorder (from the OMIM clinical synopsis or other relevant database).

However, when comparing to the OMIM PS benchmarking, the DDD tests a much narrower phenotypic spectrum of patients, heavily skewed towards ‘Abnormality of the nervous system’, ‘Abnormality of head or neck’ and ‘Abnormality of the skeletal system’ (Figure 17). These methods are developed for patients across the

entire human phenotypic spectrum so such benchmarking may be less predictive of how the methods will perform across it.

This benchmarking makes use of a logistic function to rescale patient-to-OMIM phenotype similarity scores to estimate probability of phenotype similarity, rather than assume a linear relationship. The logistic functions take the phenotype similarity between phenotypes in the same OMIM phenotypic series as the example of true similarity, and considers similarity between phenotypes and all phenotypes outside their phenotypic series as dissimilar. It is not ideal to use this simulated data, but in absence of a large dataset of patients with monogenic diagnoses and HPO deep phenotyping, the ~2,000 queries simulated queries was used as a pragmatic alternative. The development of the logistic function also assumes a constant level of significance for a particular level of $Resnik_{avg,max|sym}$ and cosine similarity across all OMIM phenotypes, which may not be the case – for one particular OMIM phenotype a certain level of patient similarity score may represent a strong match while for another OMIM phenotype this may represent a weak match. Again, a large dataset of patients with monogenic diagnoses and HPO deep phenotyping would help to resolve this, from projects such as 100,000 genomes (Genomics England, n.d.).

These benchmarking metrics tested whether methods could use patient HPO terms assigned prior to genetic diagnosis to predict the causative gene for each patient. They primarily work under the assumption that all diagnoses were correct in fully explaining the patient phenotype, something which is unable to be assessed here. Although these methods seek to predict the causative gene through assessment of all gene-phenotype relationships, in practice, the identification of a causative gene for a patient is carried out after sequencing, considering only the observed genetic

variation – there may be ~1,000 genes in the DDG2P panel used by DDD in this example but in practice this panel would be reduced to a few genes in which there is observed rare, functional variation. In the absence of available patient genetic data, one solution is to simulate exomes and spike in the pathogenic variants, which would be a viable next step in benchmarking these methods. However, the strategy undertaken here enables the assessment of the candidate gene within *all* possible developmental genes rather than the subset of genes in which there is variation, which is less realistic but a more stringent test of phenotype similarity querying methods.

There were interesting correlations between the different rank-based and score-based method performance metrics, with a negative correlation between the median and average causative gene scores for each method (Figure 20). Some methods were able to identify the correct causative gene with very high confidence in a small subset of patients but performed poorly on the rest (BOQA similarity with all reference sets and cosine similarity with the curated reference set), while other methods were able to prioritise the majority of causative genes, but with less confidence (cosine similarity with the text-mined reference sets). At one extreme, the use of BOQA resulted in the identification of the diagnostic gene with very high confidence in a small number of patients (<10%) and predicted more causal genes at rank 1, although a maximum of 23% of patients had a positive Δ. The group of methods with strong median causative gene scores were also stronger at *ranking* genes, as evidenced by the negative correlation between median gene score and gene rank, whilst the opposite is true for the former (Figure 21). This dichotomy between the two different metrics of method performance represent two valid applications of phenotype similarity methodology to candidate gene

prioritisation – some methods perform strongly on a small subset of patients, for which they can make a highly confident guess at the diagnostic gene, whereas the methods developed in this thesis demonstrate strong prioritisation performance which would be beneficial to a wider patient population.

Despite the advances shown by methods developed here, sensitivity remains relatively low for the challenging task of identifying disease-causing gene. Each combination of methods tested herein failed to assign a high rank to the causative developmental gene for a large subset of DDD patients. The top-performing method only found the causative gene within rank 10 in 23% of cases and rank 100 in 56% of cases (Figure 18). However, this is likely to reflect the dataset being enriched for patients that had a developmental disorder that was initially difficult to diagnose. More straightforward cases would therefore be depleted in this dataset (The Deciphering Developmental Disorders Study, 2014), and we would expect this methodology to perform better at solving such cases - a valid alternative application. In addition, the DDD study also identified novel gene-phenotype links, such as that of the *MED13L* gene. *MED13L* variants have previously been described in patients with congenital heart defects such as dextro-looped transposition of the great arteries (d-TGA; MIM #608808) (Muncke et al., 2003), but this study identified 8 patients with variants in *MED13L* gene that had intellectual disability but lacked congenital heart malformations (Adegbola et al., 2015).

To summarise, this chapter demonstrates the use of methods developed in the first chapter to prioritise causative genes using patient phenotypes from a diagnosed developmental disorder dataset. This also required the development of a function that rescales similarity scores to reflect the probability of phenotype similarity

which was trained on known phenotypic similarity data. Methods developed in this thesis were superior to methods developed by other groups at ranking the causative genes, although some of the previously developed methods were better at predicting diagnoses with high confidence in a small subset of the patients. Generally, the methods were poor at predicting causative genes for a large subset of the patients, although this may have been due to the nature of the dataset tested, consisting of patients that were difficult to diagnose. Application to patient data in combination with genetic data (performed in the next chapter) may help streamline method performance, because the filtered observed genetic variation will reduce the number of potential causal genes.

Chapter 4 - Use of patient phenotype similarity for genetic diagnosis in the clinic

4.1 Introduction

This chapter builds on the work done in the previous chapter where phenotype similarity methods were applied to the prioritisation of genes within a gene panel for patients with developmental disorders. Here, phenotype similarity methods were applied to the gene and variant prioritisation of exome-sequenced rare disease patients from the Guy's Hospital genetics clinic with a wide range of phenotypes. For patients with both HPO terms and a genetic diagnosis, methods developed in the previous two chapters were again compared, measuring their ability to prioritise the diagnostic gene and variant. Methods were also compared to a currently employed variant prioritisation method, PhenIX (Zemojtel et al., 2014). For patients with HPO terms and no genetic diagnosis, the methods were used to aid the identification of candidate variants.

Diagnosis of rare genetic disease has historically been an imprecise process, complicated by the differing access of clinicians and patients to diagnostic resources (e.g. diagnostic criteria), and that differing clinical specialisations of attending physicians may result in focussing on different aspects of patient phenotypes (Institute of Medicine (US) Committee on Accelerating Rare Diseases Research and Orphan Product Development, 2010). Diagnosis often requires timely referral to genetics clinics from primary care physicians that are unlikely to recognise a large number of rare diseases (Department of Health & UK Government, 2013). Prior to the widespread use of exome sequencing it was the clinical geneticist's role to evaluate the patient's clinical presentation, and then

request the appropriate genetic test to confirm the clinical diagnosis based on expert knowledge in rare genetic disease phenotypes and consultation with relevant literature (van Zelst-Stams, Scheffer, & Veltman, 2014). Whole exome and whole genome sequencing are now commonly employed because of their decline in cost – exome sequencing has been demonstrated to be more effective than gene panel testing in achieving molecular diagnoses (Neveling et al., 2013). This has changed the clinical geneticist's role – with patient genotypes from the whole exome readily available, a single variant that explains the patient's phenotype (or pair of variants in recessive disease) must be identified from within this data.

To identify a diagnostic variant, it is imperative to provide evidence that it causes phenotypes sufficiently similar to the patient (or that the variant is deleterious and present within a gene that harbours variation causing similar phenotypes). This is implicitly performed when constructing lists of genes that have been reported to cause similar phenotypes to that of the patient, which are used as a 'virtual gene panel' to filter genetic variation alongside standard variant filtering strategies (e.g. removal of synonymous and common variation). These gene lists can range from small, manually curated lists that are precisely constructed with consideration of the patient phenotype, to standard panels used for either narrow or broad disease areas (H. Lee et al., 2014; Wright et al., 2015). Gene panels assume a uniform distribution of belief across the panel that each gene could be causal – small gene lists indicate high confidence but may miss genes, while larger panels do not fully capitalise on the phenotypic similarity between the patient and known diseases – the DDG2P list comprises over 1,000 genes (Firth et al., 2009; Wright et al., 2015). Standard panels can be selectively augmented to include additional genes of interest but with the number of known monogenic gene-disease pairs now

exceeding 5,000 (Amberger et al., 2014) it is becoming less feasible to manually curate personalised virtual panels that incorporate knowledge of the entire human rare disease phenotypic spectrum.

Systematically collected patient phenotype data provides an opportunity to implement automated methodology that can utilise exhaustive knowledge of the human phenotypic spectrum for the identification of candidate genes. Methods that generate candidate gene lists tailored to the patient can offer improvements over standard virtual gene panel approaches. Firstly, automated searching across the entire human phenotypic spectrum mitigates the aforementioned issue of the growing number of known monogenic gene-disease pairs. Secondly, within a virtual panel all genes are considered equally likely to cause the disease, whereas searching across the entire human phenotypic spectrum confers the ability to score each gene based on its relevance to the patient's disease, enabling the construction of more concise and relevant gene panels.

There are several variant prioritisation tools that utilise machine-readable patient phenotypes (annotated as HPO terms) to score genes based on the belief that they could harbour variation that would cause the patient phenotype, making use of knowledge available in phenotype-gene databases and biomedical ontologies. These gene phenotype scores are then combined with variant scores of deleteriousness to produce final variant prioritisation scores. As discussed in the general introduction, eXtasy (Sifrim et al., 2013), Phen-Gen (Javed et al., 2014), Phevor (Singleton et al., 2014), PhenIX (Zemojtel et al., 2014), PHIVE (Peter N. Robinson et al., 2014) and hiPHIVE (Smedley et al., 2015) all utilise this multi-step process of defining both variant and gene phenotype scores and then combining them.

This chapter focusses on the analysis of exome sequence data of individuals that attended the genetics clinic at Guy's Hospital. Sequence data was available for 200 individuals, of whom 100 were assigned HPO terms by their clinician on Phenotips software (Girdea et al., 2013). Additionally, 95 of the 200 individuals received diagnostic variant reports – these were either class 3 (variant of unknown significance; VUS), class 4 (likely pathogenic) or class 5 (pathogenic), according to the ACMG guidelines (Ellard et al., 2017; Richards et al., 2015). In patients with both HPO terms and a likely genetic diagnosis (defined as class 4 or 5 variant report), the method comparisons from Chapters 2 and 3 were repeated to estimate their abilities to prioritise the correct diagnostic gene by querying patient HPO terms to a reference set of diseases. This also required the incorporation of a suitable variant filtering strategy. The methods developed in this thesis were compared to PhenIX (Zemojtel et al., 2014) to benchmark their ability to prioritise genetic variants based on patient phenotype – PhenIX was selected due to its strong performance against other published methods as reported in multiple reviews (Pengelly et al., 2017; Smedley & Robinson, 2015). PhenIX scores variants for both gene phenotype similarity (using the Phenomizer (Köhler et al., 2009)), and variant predicted deleteriousness using CADD (Kircher et al., 2014), PolyPhen2 (Adzhubei et al., 2013), MutationTaster (Schwarz, Cooper, Schuelke, & Seelow, 2014) and SIFT (Kumar et al., 2009), and then combines the scores to produce the final ranked prioritisation. This final prioritisation score can also be further modified to reflect whether or not the variant genotype is consistent with the suspected mode of inheritance (if any).

4.2 Materials and Methods

4.2.1 Patient details

This chapter makes use of patient genetic and phenotypic data from 200 individuals that attended the genetics clinic at Guy's hospital, with exome sequence data generated for patients within this group from 2014 onwards. All families provided informed consent for patient clinical information and sequence data to be used for research. 100 of 200 had HPO terms recorded in PhenoTips (Girdea et al., 2013), and 95 of 200 were returned diagnostic reports (following sequencing and analysis in the diagnostic lab pipeline, which involved the use of standard variant filters and virtual gene panels), of which 65 were class 4 or 5 variants (Table 10). Negative HPO term annotations were also included in HPO term annotations, but these were not incorporated into any analysis. For a further 20 patients who weren't assigned HPO terms, anonymised clinic letters were provided for the text mining of HPO terms to ascertain machine-readable clinical phenotypes.

Table 10: Number of patients from Guy's Hospital genetics clinic for whom whole exome sequence data was available, indicating numbers of patients with particular classes of diagnostic variants and whether (HPO(+)) or not (HPO(-)) their phenotypes were annotated with HPO terms.

	N	*HPO(+)*	*HPO(-)*
Total patients	200	100	100
Class 5 diagnostic report	9	7	2
Class 4 or 5 diagnostic report	65	**31**	34
Class 3, 4 or 5 diagnostic report	95	45	50
No diagnostic report	105	55	50

4.2.2 Exome sequencing analysis pipeline

For genetic reanalysis, sequences were aligned to hg19 using NovoAlign (Novocraft, 2014) and variants were called using SAMtools (H. Li et al., 2009). ANNOVAR (Wang, Li, & Hakonarson, 2010) was used for functional annotation

of variants, adding information on variant consequence, minor allele frequency in 1,000 genomes (1000 Genomes Project Consortium et al., 2015), ESP (Fu et al., 2013) and ExAC (Lek et al., 2016) datasets, and variant pathogenicity predictions using CADD (Kircher et al., 2014), SIFT (Kumar et al., 2009) and PolyPhen2 (Adzhubei et al., 2013). In-house control database (n=6,000) homozygous and heterozygous variant counts were also annotated.

4.2.3 Virtual gene panels

Primary virtual gene panels were prescribed for all patients, ranging in size from 1 to 562 genes. If a pathogenic variant was not found within the primary panel, a secondary, and possibly a tertiary, panel was applied – these panels may represent the addition of a handful of genes to augment the panel or the introduction of a completely new panel (Figure 22).

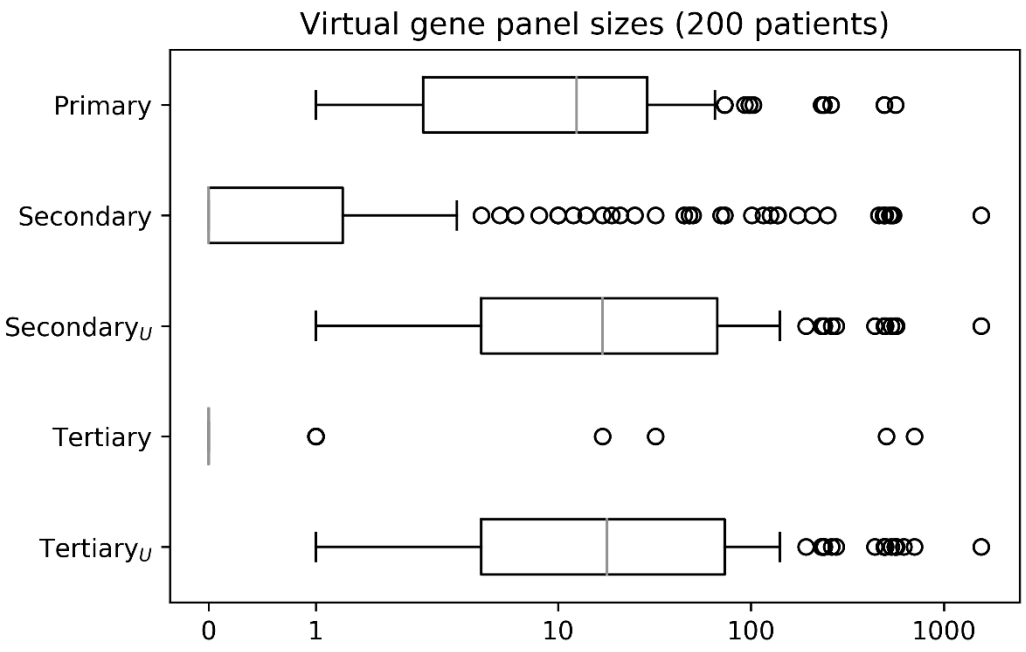


Figure 22: Primary, secondary and tertiary virtual gene panel sizes for each patient (n=200). If a secondary or tertiary panel was not used its size was considered to be zero. The subscript *U* denotes that the panel is the union of itself and panels previously prescribed – Secondary$_U$ is the union of the primary and secondary panels; Tertiary$_U$ is the union of the primary, secondary and tertiary gene panels.

4.2.4 Comparison of phenotype query methods to prioritise genes

As with analysis in the previous chapter, for the patients with both diagnostic reports and HPO terms, the different phenotype query methods were tested and their ability to prioritise the correct causative gene was assessed. Of the 31 patients with a class 4 or 5 diagnostic report, two had been reported with two variants in different genes (i.e. a digenic diagnosis), so they were removed from the analysis for simplicity. For the remaining patients their phenotype information was queried to the three different reference sets (Table 5) using both cosine similarity (Equation 9) and Resnik$_{avg,max|sym}$ (Equation 3, Equation 6, Equation 7). OMIM diseases were ranked by cosine similarity to the patient, and these phenotypic similarity scores were converted to gene scores for all OMIM-mapped genes by taking the highest phenotypic similarity score for each gene. Using this procedure, each gene was

ranked by phenotypic similarity to the patient, and methods were assessed based on the ranks of the diagnostic gene.

Table 5 (repeated): The three OMIM reference disease HPO annotation sets tested in this chapter. Only OMIM diseases captured by all annotation methods, denoted by subscript *N*, were queried against. The subscript *X* denotes those phenotypes exclusively captured by each annotation method (curated vs. text-mined). "Quantified" text-mined annotations reflect the frequency with which HPO terms are found in OMIM descriptions, whereas "unquantified" annotations have been reduced so the frequency of every term is 1.

Annotation method	**Phenotypes**	**Total annotations**	**Terms used**	**Average phenotypes per term**	**Average distance to root**
Curated	6,902	90,236	6,825	13.2	6.50
Text mining, unquantified	7,600	105,644	4,719	22.4	6.43
Text mining, quantified	7,600	230,274	4,719	22.4	6.43
Curated $_N$	6,518	88,533	6,765	13.1	6.50
Text-mined, unquantified $_N$	6,518	99,126	4,679	21.2	6.43
Text-mined, quantified $_N$	6,518	215,895	4,679	21.2	6.43
Curated $_X$	384	1,703	918	1.86	6.11
Text-mined, unquantified $_X$	1,082	6,518	1,598	4.08	6.19
Text-mined, quantified $_X$	1,082	14,379	1,598	4.08	6.19

Equation 3 (repeat): Resnik similarity between two terms (Resnik, 1999).

$$sim(t1, t2) = \quad (MICA(t1, t2))$$

Equation 6 (repeat): Best-match average term similarity between diseases Q and D

$$sim(Q \rightarrow D) = avg\left[\sum_{t1\ \epsilon\ Q} \max_{t2\ \epsilon\ D} sim(t1, t2)\right]$$

Equation 7 (repeat): Symmetrical similarity between diseases Q and D.

$$sim_{symmetric}(Q, D) = \frac{1}{2} sim(Q \rightarrow D) + \frac{1}{2} sim(D \rightarrow Q)$$

Equation 9 (repeat): Cosine similarity between vectors Q and D

$$sim(Q, D) = \cos\theta = \frac{Q \cdot D}{\|Q\| \cdot \|D\|}$$

Furthermore, using the logistic function developed in Chapter 3 (Equation 16), with the same optimised variables (Table 7), patient-OMIM phenotypic similarity scores were rescaled to reflect the probability of true similarity. These were then converted to gene scores by taking the top rescaled score for each OMIM-mapped gene, and gene scores were normalised to sum to 1.

Equation 16: Generalised logistic function used to convert phenotype similarity scores (x) to a score reflective of probability of phenotype similarity (T). Variables K, Q, B, M and v, listed for each method in Table 7, were optimised using non-linear least squares for each method in Chapter 3. K: the upper asymptote; Q: fixes the point of inflection; B: the growth rate; M: the point of maximum growth; v: asymmetry parameter.

$$T = \frac{K}{(1 + Qe^{-B(x-M)})^{1/v}}$$

Table 7: Optimised generalised logistic function variables from Equation 16 for each combination of annotation and similarity method.

Annotation method	**Similarity method**	***K***	***Q***	***B***	***M***	***v***
Curated		0.847	6.92	0.832	-0.626	0.21
Text mining, unquantified	Resnik	1	3.5	0.965	1.15	0.334
Text mining, quantified		1	1.54	1.09	3.1	0.63
Curated		0.81	0.515	6.15	0.207	0.0569
Text mining, unquantified	Cosine	1	0.628	5.46	0.753	0.598
Text mining, quantified		0.678	2.17	15.1	0.807	2.02

The list of all genes that have been established as causative of genetic disease has been termed the disease-associated genome (Zemojtel et al., 2014) – here, only genes that mapped to OMIM phenotypes within the intersection of all phenotypes annotated (Table 5) were used (n=3,303).

4.2.5 Exome variant filtering strategy

To compare different phenotype query methods to prioritise exonic variants, the following variant filtering strategy was employed:

- Removal of synonymous variants
- Removal of homozygous and compound heterozygous variants with a minor allele frequency of over 1% in either 1000 genomes (1000 Genomes Project Consortium et al., 2015), ESP (Fu et al., 2013) or ExAC (Lek et al., 2016) databases.
- Removal of remaining heterozygous variants with a minor allele frequency above 0.1%.
- Removal of homozygous and compound heterozygous variants that appear in the in-house variant database (n=6,000 individuals) more than 6 times as homozygous (0.1% of individuals) or more than 60 times as heterozygous (1% of individuals), and removal of remaining non-compound heterozygous variants appearing more than 6 times as heterozygous (0.1% of individuals) in the database.

4.2.6 Comparison of phenotype query methods to prioritise exome variants

The phenotype query methods were combined with exome variant filtering to test their ability to prioritise causative variants. After filtering, remaining variants were ranked by gene-phenotype similarity score as ranked by the aforementioned methods (4.2.4 - Comparison of phenotype query methods to prioritise genes), and compared in their ranking of the causative variant. These were also compared to the PhenIX variant prioritisation method (Zemojtel et al., 2014), which scores variants on both phenotype similarity (using the Phenomizer (Köhler et al., 2009)) and deleteriousness (incorporating CADD (Kircher et al., 2014), Polyphen2 (Adzhubei et al., 2013) and SIFT (Kumar et al., 2009) and MutationTaster (Schwarz et al., 2014) into a single metric), and combines these variant and phenotype scores into a single prioritisation score. The *variant-only*, *phenotype-*

only and *combined* scores were tested. Only variants that passed the filters listed above were included in the input VCF for PhenIX to ensure that all variant prioritisation was conducted with equivalent inputs.

Again, using the logistic function and optimised variables for each method (Equation 16, Table 7), patient-OMIM phenotypic similarity scores were rescaled to reflect the probability of similarity. These were then converted to variant scores for each filtered variant by taking the top rescaled score of their OMIM-mapped genes, and post-filtering variant scores were normalised to sum to 1 (variants were collapsed by gene during score normalisation).

4.2.7 Text mined patient annotation

For 14 patients within this dataset (for whom HPO terms were not available) a single anonymised clinic letter was made available for text mining of HPO terms (Table 11). These clinic letters were written following a visit to the genetics clinic and consisted of free form text describing the patients' clinical phenotype(s), as well as listing features present in affected family members where applicable. Two of these individuals were reported a likely pathogenic diagnostic variant and two others had been reported a VUS. To avoid sending potentially sensitive patient information to unauthorised external servers (despite anonymisation), NCBO Annotator was not used. Instead, a concept recogniser was built to identify HPO terms, synonyms and IDs, equivalent to the NCBO annotator function. Quantification was not used for text-mined patient terms, as there would be limited benefit to quantifying terms in a single letter describing a clinic visit.

Table 11: Number of exome sequenced patients from Guy's Hospital genetics clinic for whom a clinic letter was available for HPO term text mining.

	Letter(+)
Total	14
Class 5 diagnostic report	0
Class 4 or 5 diagnostic report	2
Class 3, 4 or 5 diagnostic report	4
No diagnostic report	10

4.3 Results

4.3.1 Clinic patient HPO phenotypes

When categorising patient phenotypes using the top HPO terms directly beneath ‘Phenotypic abnormality’, the clinic patients demonstrated a reasonably similar phenotypic composition to the DDD and OMIM phenotypic series (PS) query datasets used in the previous chapters, again most commonly being represented by ‘Abnormality of the nervous system’, ‘Abnormality of the skeletal system’ and ‘Abnormality of the head or neck’ clinical terms (Figure 23, Figure 24). Additionally, there was some enrichment for ‘Abnormality of prenatal development or birth’, ‘’Growth abnormality’ and ‘Abnormality of the cardiovascular system’ terms compared to the other previously benchmarked datasets.

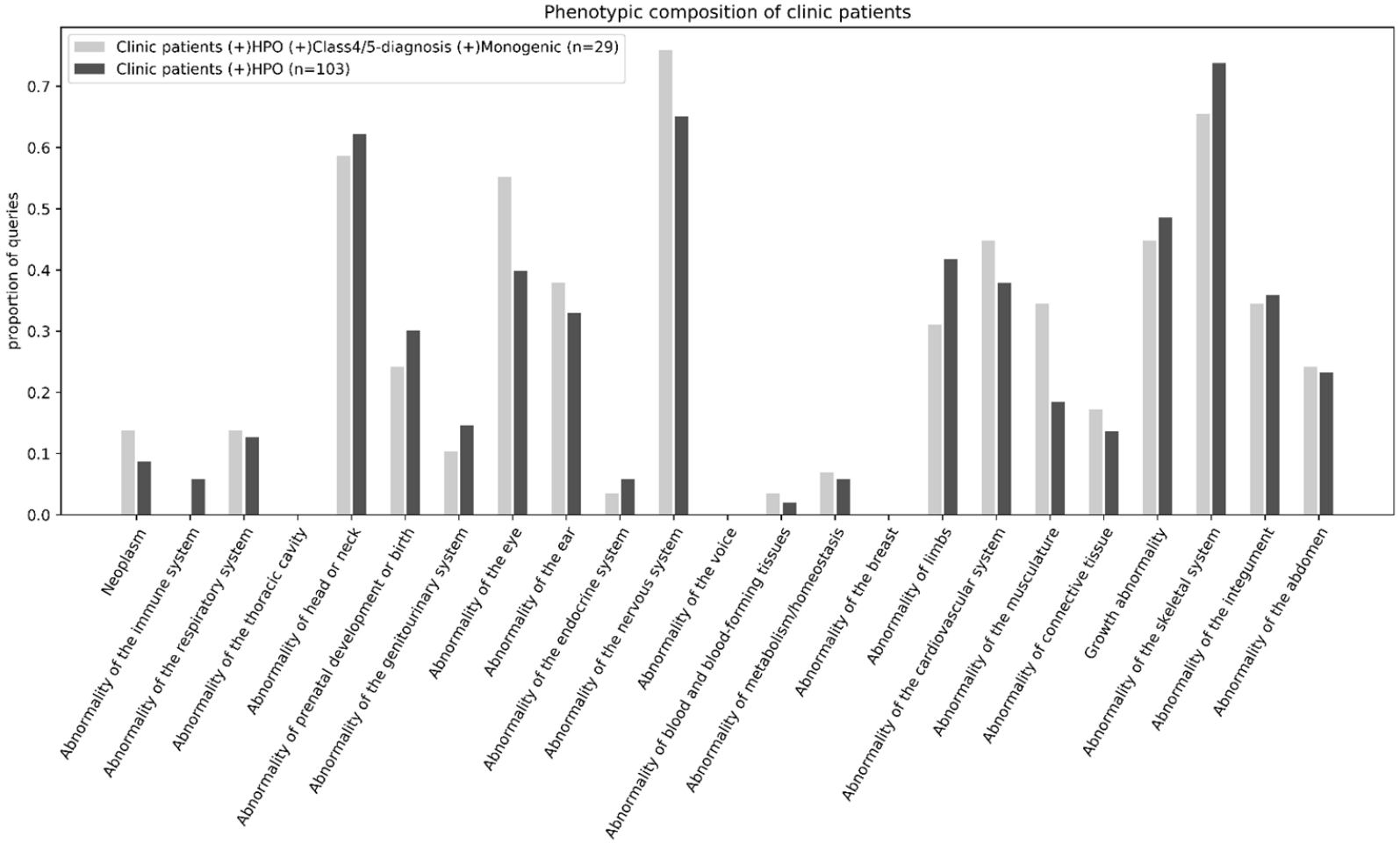


Figure 23: Range of disease phenotypes presented by the sequenced genetics clinic patients (where HPO terms were recorded), as characterised by the proportion of patients covered by each HPO term directly below 'Phenotypic abnormality' (HP:0000118).

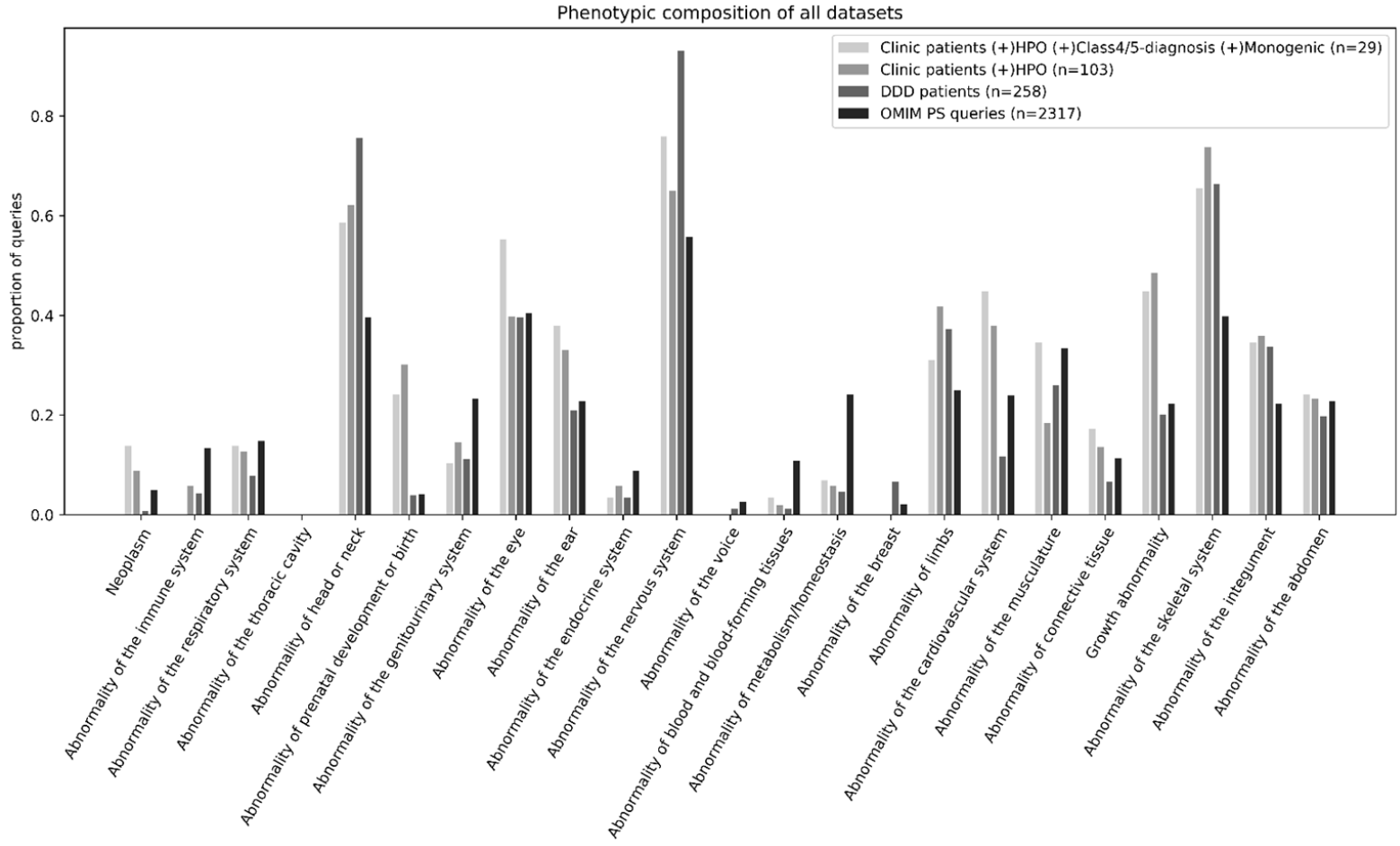
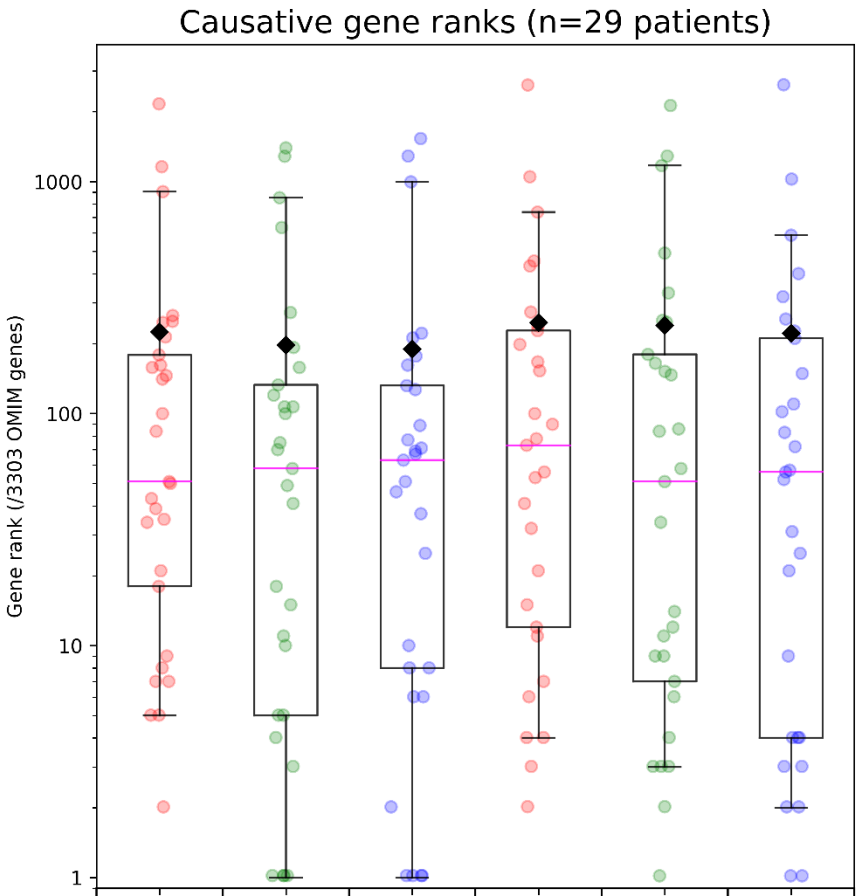


Figure 24: Range of disease phenotypes presented by the sequenced genetics clinic patients (where HPO terms were recorded) compared to the phenotypic composition of the DDD and OMIM phenotypic series datasets tested in previous chapters, as characterised by the proportion of patients covered by each HPO term directly below 'Phenotypic abnormality' (HP:0000118).

4.3.2 Method comparison on causative gene ranks

All combinations of methods for calculating gene phenotype similarity to a reference set of diseases performed similarly in prioritising genes from the OMIM disease-associated genome (n=3,303) (Figure 25).



	Res-c	Res-uq	Res-q	Cos-c	Cos-uq	Cos-q
median	51.0	58.0	63.0	73.0	51.0	56.0
mean	224.6	197.6	189.4	246.8	240.0	221.9
solved@1	0	4	4	0	1	2
solved@10	7	9	10	6	10	10
solved@100	17	18	20	18	18	18

Figure 25: Ranks of the diagnostic gene for the 29 patients with class 4/5 monogenic diagnoses and their phenotype described with HPO terms, using different combinations of reference annotations and query methods. Only phenotypes in all reference sets were queried (n=6,518) and phenotype ranks were converted to OMIM disease-associated gene ranks (n=3,303). Similarity measures: Res – $Resnik_{avg,max|sym}$; Cos – cosine similarity; Disease phenotype reference sets: c – curated; u – text-mined (unquantified); q – text-mined (quantified). Boxplot limits represent the 5th and 95th percentiles; black diamond indicates the mean rank.

The ranks of the correct gene for each of the 30 patients were compared across methods using the Wilcoxon test followed by adjustment for multiple testing under dependence, although before multiple testing correction, no two methods produced statistically significant results to one another ($P > 0.05$). Querying the unquantified text-mined reference set with cosine similarity produced the lowest median gene rank (42.5 / 3,303), while querying the quantified text-mined reference set with cosine similarity produced the lowest mean (215 / 3,303). Using $Resnik_{avg,max|sym}$ similarity with either of the text-mined sets resulted in the highest number of patients with the causative gene ranked first (4 / 29 patients). Here, there was no indication here that cosine was superior to $Resnik_{avg,max|sym}$, or that there was any benefit in quantifying the reference disease phenotype set.

4.3.3 Method comparison on causative gene scores

The rescaled scores for the correct gene for each of the 29 patients were plotted against the probability of randomly selecting a gene from the OMIM disease-associated genome (Figure 26). Scores were compared across methods using the Student's t test followed by adjustment for multiple testing under dependence. After multiple testing correction, only one comparison was statistically significant – the use of unquantified text-mined phenotypes with $Resnik_{avg,max|sym}$ similarity was significantly better than using the quantified text-mined phenotypes (with the same similarity measure) ($P = 1.11 \times 10\text{-}2$). The use of cosine similarity appeared to be superior to $Resnik_{avg,max|sym}$ in achieving a higher average fold change in correct gene probability above random (Table 12). Additionally, using cosine similarity with the curated disease phenotype reference set resulted in the greatest average causative gene probabilities for the patient group, but it had fewest patients with a probability higher than selecting an OMIM disease-associated gene at

random. The curated reference set also outperformed the other reference sets for median gene probability. Quantification of the text-mined reference set only appeared to be an improvement ahead of the unquantified version in the average gene score, and it was the unquantified version that achieved higher median score.

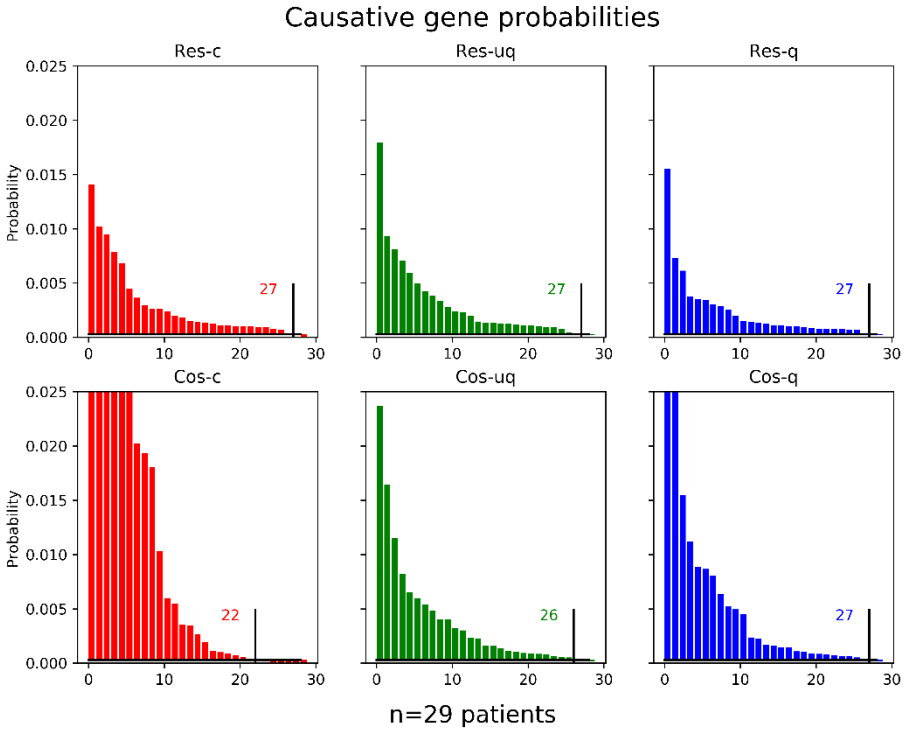


Figure 26: Probability (after logistic function rescaling) assigned to the diagnostic gene for the 29 patients with class 4/5 monogenic diagnoses and their phenotype described with HPO terms, using different combinations of reference annotations and query methods. Probability was plot against a baseline of selecting an OMIM disease-associated gene at random (1/3303). Similarity measures: Res – $Resnik_{avg,max|sym}$; Cos – cosine similarity; Disease phenotype reference sets: c – curated; u – text-mined (unquantified); q – text-mined (quantified). Number of patients for which probability is higher than randomly selecting a OMIM disease-associated gene is indicated on the plot.

Table 12: Metrics for score-based analysis (after logistic function rescaling) of different phenotype annotation and similarity methods used to identify the causative genes for diagnosed genetics clinic patients (from Figure 26) Δ = probability – prior; Fold change = Δ/prior.

Annotation method	Similarity method	Average Δ	Average fold change	Median Δ	Median fold change	n(Δ > 0)
Curated	Resnik	2.66E-03	8.781	1.14E-03	3.773	27
Text-mined, unquantified		2.78E-03	9.177	1.06E-03	3.495	27
Text-mined, quantified		2.02E-03	6.669	8.40E-04	2.775	27
Curated	Cosine	1.35E-02	44.429	2.32E-03	7.677	22
Text-mined, unquantified		3.64E-03	12.016	1.34E-03	4.417	26
Text-mined, quantified		5.76E-03	19.01	1.32E-03	4.367	27

4.3.4 Exome sequencing

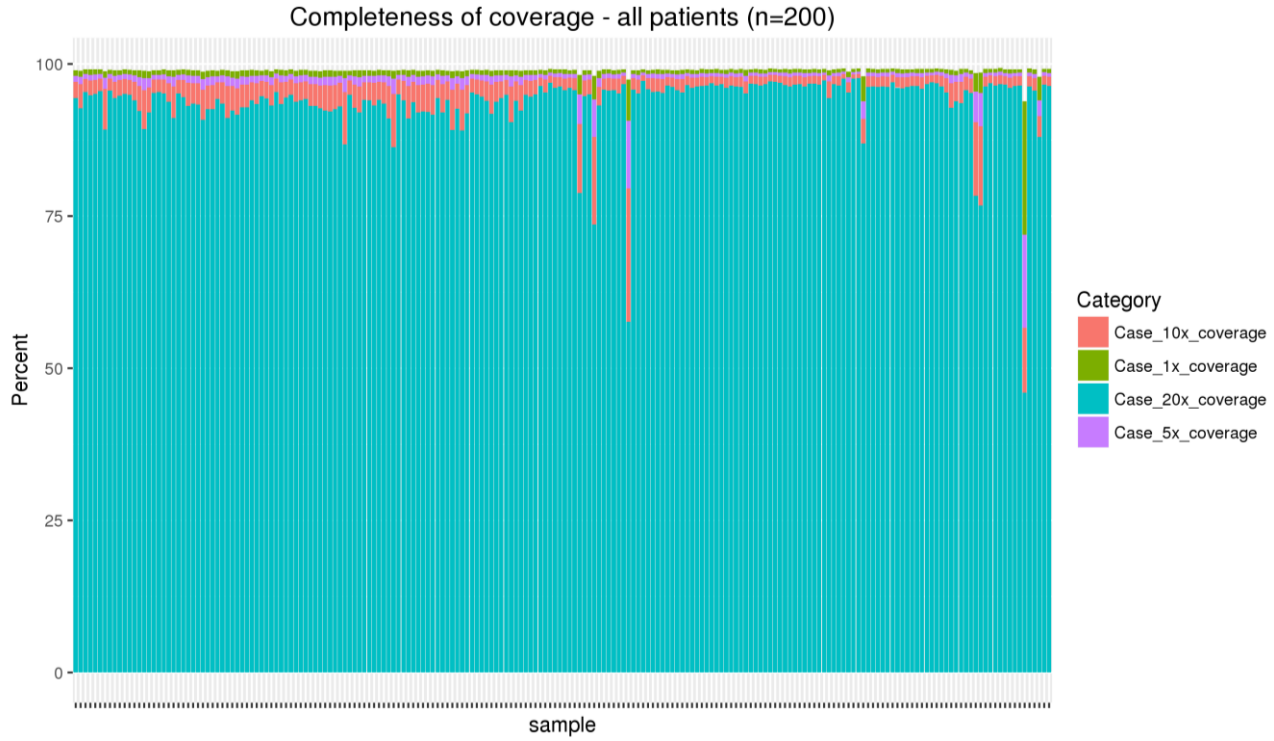


Figure 27: Exome coverage statistics for all sequenced genetics clinic patients (n=200). Percentage coverage (y-axis) given for each individual (x-axis) at 20X, 10X, 5X and 1X.

Sequence coverage was assessed after exome sequencing (Figure 27, Figure 28). 98.5% of individuals had 75% coverage at 20X (including all diagnosed individuals) and 93% had 90% coverage at 20X.

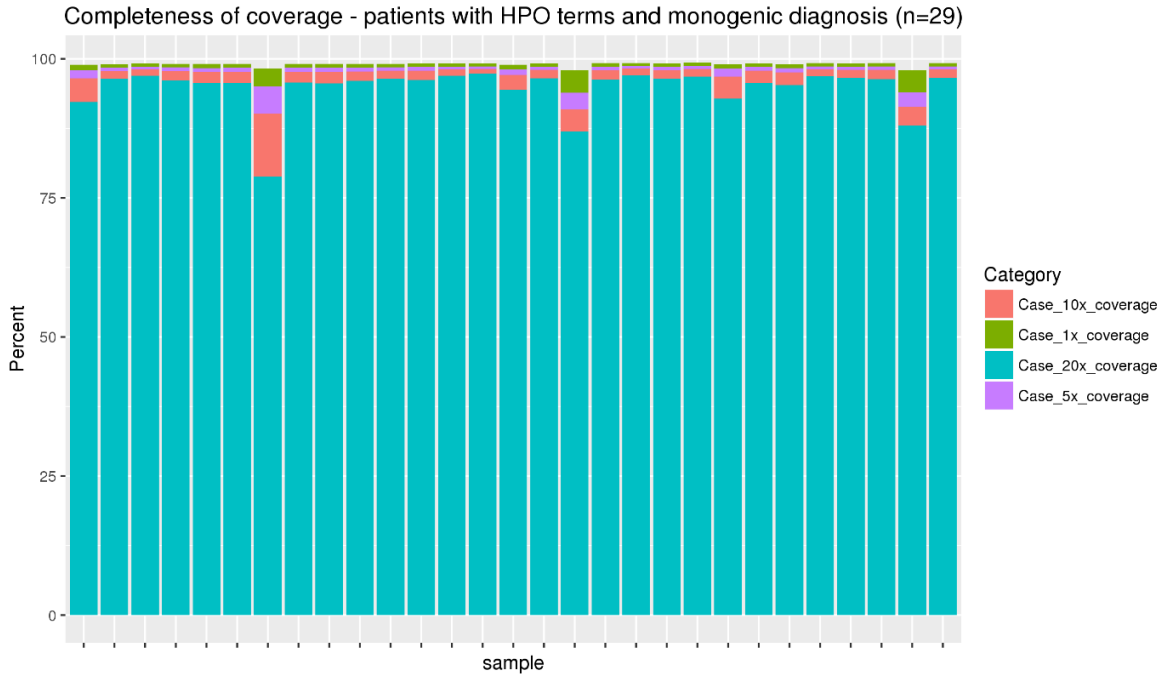


Figure 28: Exome coverage statistics for sequenced genetics clinic patients with monogenic diagnoses and assigned HPO terms (n=29). Percentage coverage (y-axis) given for each individual (x-axis) at 20X, 10X, 5X and 1X.

4.3.5 Exome variant filtering

Filters were applied to the annotated exome variants leaving a median of 316 variants in all HPO-annotated patients, of which there were a median of 71 variants in OMIM disease-associated genes (Figure 29).

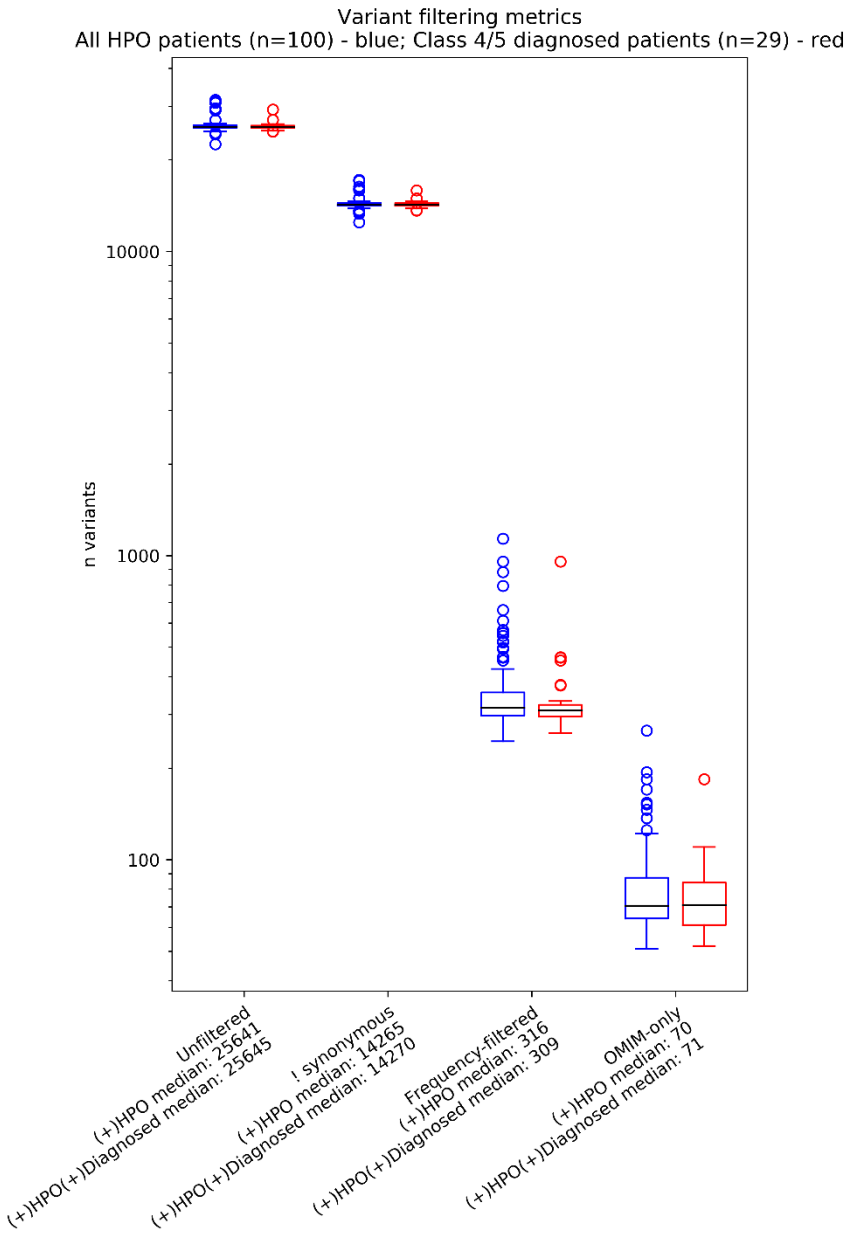


Figure 29: Exome variant filtering statistics, for both the entire HPO-annotated patient group (blue; n=100) and the HPO-annotated group with monogenic class 4/5 diagnoses (red; n=29). Filtering was applied sequentially, applying steps stated on the *x*-axis from left to right. Patients with low coverage sequencing were not included in this plot.

4.3.6 Identification of diagnostic variants

Of the 29 patients with HPO terms and monogenic diagnoses, the diagnostic variant was captured following the exome filtering strategy in 27 cases. The two variants that weren't captured are as follows:

Variant (i) *MED12:NM_005120:exon42:c.6256_6258del:p.2086_2086del*

A heterozygous nonframeshift deletion in exon 42 of MED12 in a patient (HPO terms in Appendix 1) with good phenotypic matches to the three autosomal dominant *MED12* disorders, Opitz-Kaveggia syndrome (OMIM:305450), Lujan-Fryns syndrome (OMIM:309520) and X-linked OHDO syndrome (OMIM:300895).

The reported nonframeshift mutation was not represented in any of the public databases (1,000 genomes, ESP or ExAC), and had no pathogenicity prediction scores available. It was observed 8 times as heterozygous in the in-house database, and therefore was rejected by the variant filtering strategy. The variant also exists in ClinVar with recently reported conflicting interpretations of pathogenicity (twice as a VUS, once as likely benign). Although there is a possibility that this is the true causative variant due to the phenotypic match, this may a misclassification of a variant of unknown significance, as the ClinVar reports were unlikely to be available at the time of diagnosis.

Variant (ii) *CREBBP:NM_004380:exon14:c.C2678T:p.S893L*

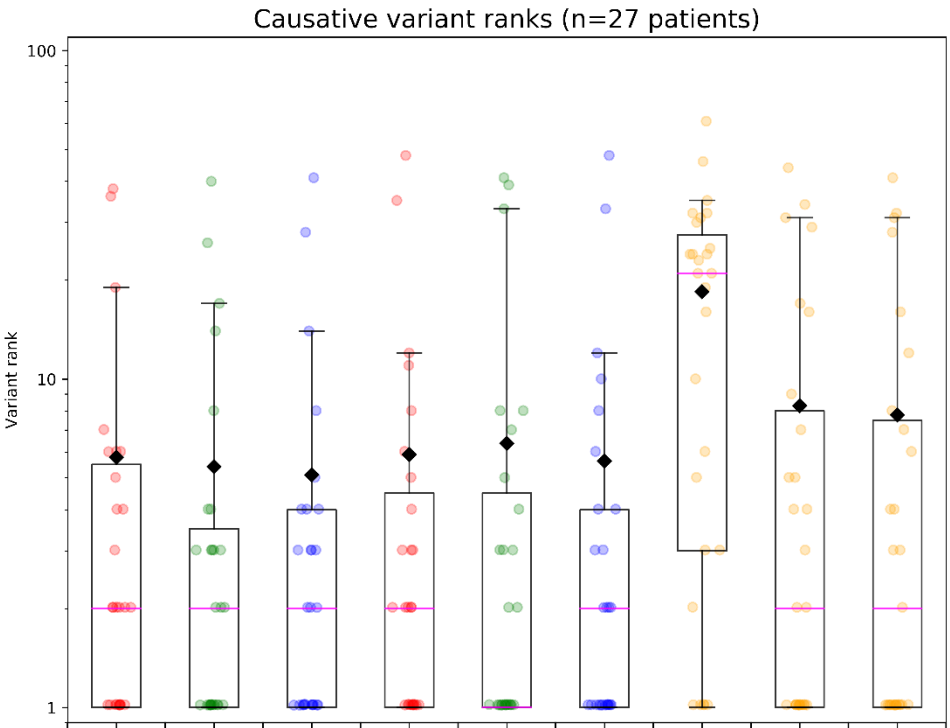
A heterozygous nonsynonymous SNV in exon 14 of *CREBBP* in a patient (HPO terms in Appendix 1) with a good phenotypic match to Rubinstein-Taybi syndrome (OMIM:180849), an autosomal dominant disorder. The variant had a CADD score indicative of pathogenicity (21.0), although SIFT and PolyPhen2 scores were less

indicative (0.21 and 0.357 respectively). The variant was found at a MAF slightly below 0.1% in ExAC and 1,000 genomes, but it was found at 0.14% in ESP and therefore was rejected by the filtering strategy used. The variant was also found in ClinVar reported as either ‘benign’ or ‘likely benign’, which again may be a misclassification made before the reports on ClinVar, despite the strong gene-phenotype match with the patient and a CADD score that is indicative of deleteriousness.

4.3.7 Method comparison on causative variant ranks

The phenotype-aware methods of variant prioritisation (all methods but PhenIX-V, which only considers variant score) all performed reasonably similarly in ranking causative variants (Figure 30). The ranks of the correct gene for each of the 27 patients were compared across methods using the Wilcoxon test followed by adjustment for multiple testing under dependence. No method performed significantly better than any other after multiple testing correction ($P > 0.05$). All phenotype-aware methods showed an improvement over the PhenIX variant-only score prioritisation, although this difference did not survive multiple testing correction. The optimal method for achieving a top median variant rank was cosine similarity combined with the unquantified text-mined reference set, which also ranked the highest number of causative variants in first place (14 of 27). There was no indication that cosine similarity outperformed $\text{Resnik}_{\text{avg,max|sym}}$ with these benchmarking metrics, nor was there any indication that there was any benefit in quantifying reference phenotype data. Methods developed in this work performed similarly to the PhenIX phenotype and phenotype-variant combined prioritisation. There was no indication that incorporating the variant score into the combined

score resulted in greater performance of PhenIX (apart from a fractionally better average rank).



	Res-c	Res-uq	Res-q	Cos-c	Cos-uq	Cos-q	PhenIX-V	PhenIX-P	PhenIX-C
median	2.0	2.0	2.0	2.0	1.0	2.0	21.0	2.0	2.0
mean	5.8	5.4	5.1	5.9	6.4	5.6	18.4	8.3	7.8
solved@1	10	12	12	11	14	13	5	12	13
solved@10	24	23	24	23	24	24	11	21	21
solved@100	27	27	27	27	27	27	27	27	27

Figure 30: Ranks of the diagnostic variant (after filtering) for the 27 patients with class 4/5 monogenic diagnoses and their phenotype described with HPO terms, using different combinations of reference annotations and query methods. Only phenotypes in all reference sets were queried (n=6,518) and variants were ranked by gene phenotype similarity to the patient. Similarity measures: Res – $Resnik_{avg,max|sym}$; Cos – cosine similarity; Disease phenotype reference sets: c – curated; u – text-mined (unquantified); q – text-mined (quantified); PhenIX prioritisation settings: PhenIX-V – PhenIX variant pathogenicity score; PhenIX-P – PhenIX variant gene phenotype score; PhenIX-C – PhenIX combined score. Boxplot limits represent the 5th and 95th percentiles; black diamond indicates the mean rank.

4.3.8 Method comparison on causative variant probabilities

The scores for the diagnostic variant for each of the 27 remaining patients were plot against the baseline probability of selecting variants from a random gene as causative, using the median number of post-filtering variants: 62 (Figure 31). Here, the post-filtering variants were collapsed by gene (hence the discrepancy with

Figure 29). Scores were compared across methods using the Student's *t* test followed by adjustment for multiple testing under dependence. Using cosine similarity combined with the curated reference set was significantly better than all other methods but cosine similarity with quantified text-mined phenotypes, although it had fewest patients for whom the probability of the diagnostic variant was higher than randomly selecting a post-filtering OMIM variant. The curated reference set also outperformed the other reference sets for median gene probability (Table 13). Using cosine similarity with the quantified text-mined reference phenotypes was only significantly better in scoring causative variants compared to using $\text{Resnik}_{\text{avg,max|sym}}$ with the same reference phenotypes. There was some indication that cosine similarity was superior to Resnik similarity for these benchmarking metrics, which was statistically significant for the curated and quantified text-mined reference sets (in both average and median causative variant score).

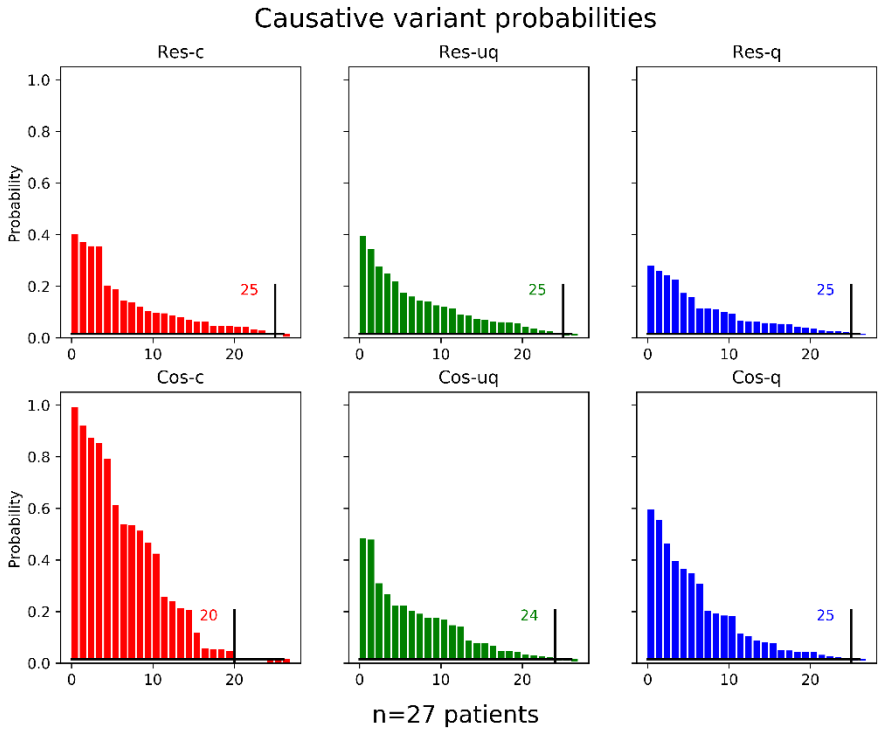


Figure 31: Probability (after logistic function rescaling) assigned to the diagnostic variant for the 27 patients with class 4/5 monogenic diagnoses and their phenotype described with HPO terms, using different combinations of reference annotations and query methods. Probability was plot using the median number of variants remaining after filtering (whose gene was mapped to an OMIM phenotype) as a baseline of selecting a variant at random (1/62). Similarity measures: Res – $Resnik_{avg,max|sym}$; Cos – cosine similarity; Disease phenotype reference sets: c – curated; u – text-mined (unquantified); q – text-mined (quantified). Number of patients for which probability is higher than randomly selecting a variant is indicated on the plot.

Table 13: Metrics for score-based analysis (after logistic function rescaling) of different phenotype annotation and similarity methods used to identify the causative variants for the diagnosed genetics clinic patients (from Figure 31). Δ = probability – prior; Fold change = Δ/prior.

Annotation method	Similarity method	Average Δ	Average fold change	Median Δ	Median fold change	n(Δ > 0)
Curated	Resnik	1.05E-01	6.482	6.37E-02	3.95	25
Text-mined, unquantified		1.02E-01	6.333	6.95E-02	4.312	25
Text-mined, quantified		7.76E-02	4.81	4.58E-02	2.839	25
Curated	Cosine	3.11E-01	19.268	1.97E-01	12.191	20
Text-mined, unquantified		1.24E-01	7.69	7.04E-02	4.366	24
Text-mined, quantified		1.55E-01	9.606	7.15E-02	4.431	25

4.3.9 Candidate variants in patients without diagnoses

Of the 55 patients without diagnostic reports and with HPO terms describing their phenotypes, filtered variants were prioritised according to the results of querying the patient phenotype to the quantified text-mined OMIM reference set using cosine similarity (and assigned a variant score using the logistic function). Variants were evaluated with respect to patient phenotype, their zygosity was compared to the reported mode(s) of inheritance of gene OMIM phenotypes, and pathogenicity scores from CADD, SIFT and PolyPhen2. This process yielded interesting candidate variants in single genes in 7 patients (listed in Table 14-Table 20).

Table 14: Patient I – *LRP5* variant(s) identified after querying patient HPO terms to the quantified text-mined OMIM disease phenotype reference set using cosine similarity. *Var[iation] gene rank* is the rank of the gene within all OMIM disease-associated genes for which there was observed variation that passed the exome variant filters. *Variant probability* is calculated using the same procedure used for benchmarking methods (4.2.6 - Comparison of phenotype query methods to prioritise exome variants). *OMIM phenotypes* are listed in order of similarity to the patient HPO terms, along with their reported mode of inheritance. *Minor allele frequencies (MAF)*, *CADD*, *SIFT* and *Polyphen2* scores, as well as *ClinVar* assessments are listed for each variant (separated by `//` where multiple variants are identified).

Patient	P(i)
HPO terms	HP:0009778 - Short thumb HP:0001156 - Brachydactyly syndrome HP:0010239 - Aplasia of the middle phalanx of the hand HP:0005807 - Absent distal phalanges
Gene	LRP5
Variant(s)	Het exon1:c.58_60del:p.20_20del Het exon18:c.C3864A:p.D1288E
Variation gene rank	8
Variant probability	0.01759
OMIM phenotypes	OMIM:607636 VAN BUCHEM DISEASE, TYPE 2 (AD) OMIM:144750 ENDOSTEAL HYPEROSTOSIS, (AD) OMIM:601884 BONE MINERAL DENSITY QUANTITATIVE TRAIT LOCUS 1; BMND1 (AD) OMIM:166710 OSTEOPOROSIS (AD) OMIM:607634 OSTEOPETROSIS, AUTOSOMAL DOMINANT 1; OPTA1 (AD) OMIM:259770 OSTEOPOROSIS-PSEUDOGLIOMA SYNDROME; OPPG (AR) OMIM:601813 EXUDATIVE VITREORETINOPATHY 4; EVR4 (AD/AR)
ESP MAF	0 // 0
ExAC MAF	0 // 0
1000 genomes MAF	0 // 0
CADD	NA // 31
SIFT	NA // 0
PolyPhen2	NA // 0.999
ClinVar	Likely benign // Absent

Table 15: Patient II – *ZNF423* variant(s) identified after querying patient HPO terms to the quantified text-mined OMIM disease phenotype reference set using cosine similarity. See Table 14 legend for a breakdown of the headings.

Patient	P(ii)
HPO terms	HP:0000104 - Renal agenesis HP:0000151 - Aplasia of the uterus
Gene	ZNF423
Variant(s)	Het exon4:c.G1471A:p.D491N
Variation gene rank	4
Probability	0.04597
OMIM phenotypes	OMIM:614844 NEPHRONOPHTHISIS 14; NPHP14 (AD/AR)
ESP MAF	0.0002
ExAC MAF	0.0002
1000 genomes MAF	0.0001999
CADD	23.9
SIFT	NA
PolyPhen2	0.901
ClinVar	1 report of VUS for same disorder

Table 16: Patient III – *COL4A3BP* variant(s) identified after querying patient HPO terms to the quantified text-mined OMIM disease phenotype reference set using cosine similarity. See Table 14 legend for a breakdown of the headings.

Patient	P(iii)
HPO terms	HP:0000179 - Thick lower lip vermilion HP:0001654 - Abnormality of the heart valves HP:0001169 - Broad palm HP:0001263 - Global developmental delay HP:0000426 - Prominent nasal bridge HP:0000455 - Broad nasal tip HP:0000637 - Long palpebral fissure HP:0000316 - Hypertelorism HP:0000470 - Short neck HP:0000914 - Shield chest HP:0000736 - Short attention span
Gene	COL4A3BP
Variant(s)	Het exon2:c.129_130insACGGCG:p.T44delinsTAT
Var. gene rank	4
Probability	0.05756
OMIM phenotypes	OMIM:616351 MENTAL RETARDATION, AUTOSOMAL DOMINANT 34; MRD34 (AD)
ESP MAF	0
ExAC MAF	0
1000 genomes MAF	0
CADD	NA
SIFT	NA
PolyPhen2	NA
ClinVar	Absent

Table 17: Patient IV – *FAT4* variant(s) identified after querying patient HPO terms to the quantified text-mined OMIM disease phenotype reference set using cosine similarity. See Table 14 legend for a breakdown of the headings.

Patient	P(iv)
HPO terms	HP:0009778 - Short thumb HP:0005321 - Mandibulofacial dysostosis HP:0000256 - Macrocephaly HP:0001680 - Coarctation of aorta
Gene	FAT4
Variant(s)	Het exon1:c.G185C:p.G62A Het exon1:c.C4250A:p.P1417H
Var. gene rank	7
Probability	0.02356
OMIM phenotypes	OMIM:615546 VAN MALDERGEM SYNDROME 2; VMLDS2 (AR) OMIM:616006 HENNEKAM LYMPHANGIECTASIA-LYMPHEDEMA SYNDROME 2; HKLLS2 (AR)
ESP MAF	0.0011 // 0.0002
ExAC MAF	0.0011 // 4.97E-05
1000 genomes MAF	0.0003994 // 0.0003994
CADD	25.4 // 27
SIFT	0.35 // 0.5
PolyPhen2	0.997 // 0.979
ClinVar	Absent // VUS

Table 18: Patient V – *DNM2* variant(s) identified after querying patient HPO terms to the quantified text-mined OMIM disease phenotype reference set using cosine similarity. See Table 14 legend for a breakdown of the headings.

Patient	P(v)
HPO terms	HP:0000220 - Velopharyngeal insufficiency HP:0001199 - Triphalangeal thumb HP:0003508 - Proportionate short stature HP:0001643 - Patent ductus arteriosus HP:0002020 - Gastroesophageal reflux HP:0000324 - Facial asymmetry HP:0000405 - Conductive hearing impairment
Gene	DNM2
Variant(s)	Het exon6:c.G797A:p.R266Q
Var. gene rank	34
Probability	0.01261
OMIM phenotypes	OMIM:160150 MYOPATHY, CENTRONUCLEAR, 1; CNM1 (AD) OMIM:615368 LETHAL CONGENITAL CONTRACTURE SYNDROME 5; LCCS5 (AR) OMIM:606482 CHARCOT-MARIE-TOOTH DISEASE, DOMINANT INTERMEDIATE B; CMTDIB (AD)
ESP MAF	0
ExAC MAF	0
1000 genomes MAF	8.318E-06
CADD	34
SIFT	0.02
PolyPhen2	0.991
ClinVar	Absent

Table 19: Patient VI – *NOD2* variant(s) identified after querying patient HPO terms to the quantified text-mined OMIM disease phenotype reference set using cosine similarity. See Table 14 legend for a breakdown of the headings.

Patient	P(vi)
HPO terms	HP:0000988 - Skin rash HP:0001382 - Joint hypermobility HP:0012432 - Chronic fatigue
Gene	NOD2
Variant(s)	Het exon4:c.G2332A:p.E778K
Var. gene rank	1
Probability	0.1968
OMIM phenotypes	OMIM:186580 BLAU SYNDROME; BLAUS (AD)
ESP MAF	0.0002
ExAC MAF	0.0002
1000 genomes MAF	0.0001997
CADD	25.8
SIFT	0.02
PolyPhen2	0.999
ClinVar	VUS

Table 20: Patient VII – *COL4A3* variant(s) identified after querying patient HPO terms to the quantified text-mined OMIM disease phenotype reference set using cosine similarity. See Table 14 legend for a breakdown of the headings.

Patient	P(vii)
HPO terms	HP:0000077 - Abnormality of the kidney HP:0011604 - Aortopulmonary window HP:0010880 - Increased nuchal translucency HP:0002251 - Aganglionic megacolon HP:0009601 - Aplasia/Hypoplasia of the thumb HP:0000598 - Abnormality of the ear
Gene	COL4A3
Variant(s)	Het exon32:c.G2647A:p.G883R
Var. gene rank	17
Probability	0.01794
OMIM phenotypes	OMIM:203780 ALPORT SYNDROME (AR) OMIM:104200 ALPORT SYNDROME (AD) OMIM:141200 HEMATURIA, BENIGN FAMILIAL; BFH (AD)
ESP MAF	0
ExAC MAF	0
1000 genomes MAF	0
CADD	23.3
SIFT	0
PolyPhen2	1
ClinVar	Absent

4.3.10 Clinic letter text mining

Anonymised clinic letters were provided for 14 of the sequenced patients, which were text-mined for HPO terms. These were letters describing a single visit to the genetics clinic, between 359 and 1,068 words (Figure 32). Patient phenotypes were again queried to the quantified text-mined phenotype reference set using cosine similarity as with the undiagnosed patients described in the previous section (4.3.9).

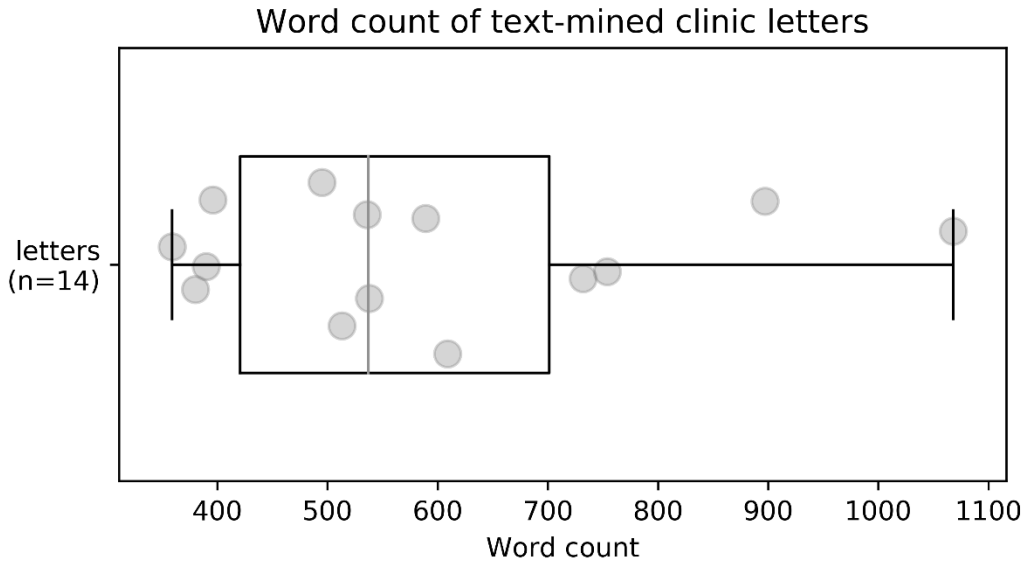


Figure 32: Word counts of the 14 anonymised clinic letters available for text mined HPO phenotype annotation.

Class 4 diagnostic variants (n=2)

Two of the sequenced patients for whom anonymised clinic letters were available for text mining had class 4 diagnostic variants. Both of these were heterozygous variants: one in exon 52 *PIEZO2* (Table 21) and one in exon 6 *TP63* (Table 22). Following variant filtering (4.2.5) and gene phenotype ranking, both causative variants were ranked first out of the remaining variants (n = 64 and 57 respectively).

Table 21: PIEZO2 variant diagnosis captured at the top variant rank after using text mining to identify query HPO terms. See Table 14 legend for a breakdown of the headings.

Patient	TM(i)
HPO terms identified	HP:0012825 – Mild HP:0009821 - Forearm undergrowth HP:0000260 - Wide anterior fontanel HP:0002804 - Arthrogryposis multiplex congenita HP:0012385 - Camptodactyly HP:0000175 - Cleft palate HP:0000006 - Autosomal dominant inheritance HP:0004322 - Short stature HP:0004324 - Increased body weight HP:0001371 - Flexion contracture HP:0000431 - Wide nasal bridge HP:0000316 - Hypertelorism HP:0001883 - Talipes HP:0000347 – Micrognathia
Gene	PIEZO2
Variant(s)	Het exon52:c.G8045A:p.G2682E
Var. gene rank	1
Probability	0.5193
OMIM phenotypes	OMIM:114300 ARTHROGRYPOSIS, DISTAL, TYPE 3; DA3 (AD) OMIM:248700 MARDEN-WALKER SYNDROME; MWKS (AD) OMIM:108145 ARTHROGRYPOSIS, DISTAL, TYPE 5; DA5 (AD)
ESP MAF	0
ExAC MAF	0
1000 genomes MAF	0
CADD	29.6
SIFT	0.0
PolyPhen2	0.997
ClinVar	Absent

Table 22: TP63 variant diagnosis captured at the top variant rank after using text mining to identify query HPO terms. See Table 14 legend for a breakdown of the headings.

Patient	TM(ii)
HPO terms identified	HP:0000668 – Hypodontia HP:0000975 - Hyperhidrosis HP:0000164 - Abnormality of the teeth HP:0001649 - Tachycardia HP:0012825 - Mild HP:0100874 - Thick hair HP:0001263 - Global developmental delay HP:0000691 - Microdontia HP:0000968 - Ectodermal dysplasia HP:0001903 - Anemia HP:0000964 – Eczema
Gene	TP63
Variant(s)	Het exon6:c.G797T:p.R266L
Var. gene rank	1
Probability	0.2002
OMIM phenotypes	OMIM:604292 ECTRODACTYLY, ECTODERMAL DYSPLASIA, AND CLEFT LIP/PALATE SYNDROME 3; EEC3 (AD) OMIM:103285 ADULT SYNDROME (AD) OMIM:129400 RAPP-HODGKIN SYNDROME; RHS (AD) OMIM:106260 ANKYLOBLEPHARON-ECTODERMAL DEFECTS-CLEFT LIP/PALATE (AD) OMIM:603543 LIMB-MAMMARY SYNDROME; LMS (AD) OMIM:605289 SPLIT-HAND/FOOT MALFORMATION 4; SHFM4 (AD)
ESP MAF	0
ExAC MAF	0
1000 genomes MAF	0
CADD	28.7
SIFT	NA
PolyPhen2	0.947
ClinVar	Absent

Interesting variants in an undiagnosed patient

An interesting set of variants was identified in the ADAMTS10 gene (Table 23), which had the top gene phenotype similarity ranking amongst the post-filtering variants. The ADAMTS10 gene was not contained within any virtual gene panels

used for the patient. Weill-Marchesani syndrome was also the top phenotypic match to the patient. All variants were not observed in the ESP, ExAC and 1000 genomes datasets. Two of the variants were frameshifts, while the splicing variant in exon 14 had a CADD score of 19.84 (just outside the top 1% of all variant scores for deleteriousness). Further investigation is necessary including sequencing of parental samples will reveal if these variants exist in *cis* or *trans*.

Table 23: Interesting variants in ADAMTS10 at the top variant rank following use of text mining to identify query HPO terms. See Table 14 legend for a breakdown of the headings.

Patient	TM(iii)
HPO terms	HP:0012837 – Generalized HP:0012531 - Pain HP:0001156 - Brachydactyly syndrome HP:0004322 - Short stature HP:0001642 - Pulmonic stenosis HP:0001387 - Joint stiffness
Gene	ADAMTS10
Variant(s)	Het ADAMTS10:exon4:c.94_98del:p.F32fs Het ADAMTS10:exon4:c.176_177insAGT:p.P59delinsPV Het ADAMTS10:exon4:c.204dupG:p.T68fs Het ADAMTS10:exon14:c.1480-9G>A
Var. gene rank	1
Probability	0.4638
OMIM phenotypes	OMIM:277600 WEILL-MARCHESANI SYNDROME 1; WMS1 (AR)
ESP MAF	0 (all)
ExAC MAF	0 (all)
1000 genomes MAF	0 (all)
CADD	19.84 for the exon 14 variant (others: NA)
SIFT	NA (all)
PolyPhen2	NA (all)
ClinVar	Absent (all)

4.4 Discussion

The clinical genetics patient dataset used in this study was reasonably similar in phenotypic constitution to the OMIM phenotypic series and DDD datasets of the previous two chapters (Figure 24) (with some enrichments for a few additional root HPO terms), so similar benchmarking results were expected. However, only 50% of the 100 patients had their phenotypes described with HPO terms (Table 10), and only 47.5% of all patients had been issued reports with diagnostic variants. This left 22.5% (45 of 200) with both HPO terms and a diagnostic report – however, to accurately benchmark these methods, class 3 diagnoses (variants of unknown significance) were discarded, leaving only clear likely pathogenic diagnostic variants (31 of 200; 15.5%). A further two patients were discarded due to having diagnostic reports in multiple genes, and the diagnostic variants for an additional two patients were not identified using the variant filtering strategy employed. With only 27 patients available for benchmarking there was lower statistical power to make conclusions on the differences between gene/variant prioritisation methods compared to the DDD and OMIM PS datasets of previous chapters, which utilised 258 and 2317 queries respectively.

There were also limitations in the comparison between the utility of phenotype-based gene prioritisation techniques and the use of virtual gene panels in this dataset. Firstly, it was unclear to what extent the initial process of identifying diagnostic variants was performed independently of assigned HPO (or other phenotypic) terms, or if tools that utilise machine-readable phenotype tools were used to help curate the virtual gene panels. PhenoTips contains an in-built differential diagnosis recommender (Girdea et al., 2013), and although the use of tools such as the Phenomizer (Köhler et al., 2009) are not standard practice in the

clinic, this does not preclude the possibility of individual clinicians making use of online resources. Secondly, as displayed by the varied sizes of gene panels used for each patient (and that in some instances secondary and tertiary panels were used) (Figure 22), the application of virtual gene panels in these cases represents a wide range of levels of confidence in the panel. 73 of 200 primary panels contained five or fewer genes and 35 comprised a single gene, demonstrating very high confidence in such genes, whilst some secondary panels (only applied if the primary panel yielded no results) contained more than 1,000 genes. Although it is implicit in the design of these methods that they should perform well on patients for whom there is less difficulty in identifying the diagnostic variant, a key reason for their development is for more 'difficult' cases, as there are now many documented monogenic causative gene-phenotype relationships which are not necessarily all readily known. The DDD dataset consisted of patients for whom diagnosis using standard gene testing was unsuccessful and therefore was a good representation of patients for whom this methodology would benefit the most. Differential method performance between these two datasets perhaps reflects this – in the best case the causative gene was identified within rank 10 for 23% of DDD patients (Figure 18), whereas the same was true for 34% of the clinic patients tested in this chapter (Figure 25), despite only genes within the DDG2P panel being tested for the DDD patients, which had roughly half the number of genes as the OMIM disease-associated genome. It would also be useful to test machine-readable phenotype similarity methods on different datasets that cover separate disease areas (or to have a large enough sample size to be able to separate the dataset) – this may reveal phenotypic areas for which there is greater need for ontology development, as well

as disease areas where there are a lack of documented phenotype-gene relationships which need further exploration.

The two diagnostic variants that were excluded by the filtering strategy employed here were heterozygous variants in *MED12* and *CREBBP* genes – one did not pass the 0.1% MAF filter in the public variant databases, and the other did not pass the 0.1% heterozygous filter in the in-house database. Both variants were in genes with good phenotypic matches to their respective patients, although both variants were annotated as either VUS or benign by multiple reporters on ClinVar. Considering the frequency at which these variants were found, it is possible that these are variants of unknown significance rather than likely pathogenic diagnoses.

Although the benchmarked sample size tested here was too small to draw many statistically significant conclusions regarding the differential performance of methods in prioritising variants and genes, there was some concordance observed with the DDD benchmarking, along with some instances of discordance. As observed in the DDD patients, using cosine similarity to measure similarity between patient and reference set was advantageous over $Resnik_{avg,max|sym}$ in assigning high probabilities to causative genes within the disease-associated genome (Figure 26, Table 12) and causative variants (Figure 31, Table 13). Cosine similarity also appeared to have an advantage over $Resnik_{avg,max|sym}$ in ranking causative variants (Figure 30) but, discordant to observations from the DDD dataset, it showed no advantage over Resnik in for ranking causative genes within the disease-associated genome (Figure 25). Also concordant with the DDD results, using cosine similarity with the curated disease phenotype reference set resulted in the greatest average causative gene probabilities for the patient group, as well as the fewest diagnostic genes with a probability higher than selecting an OMIM

disease-associated gene at random. Unlike the DDD benchmarking, using the curated reference set also outperformed the other reference sets for median gene probability. The quantified text-mined phenotype reference set showed some benefit, but only compare to the curated reference set in ranking causative variants (Figure 30). However, unlike the DDD benchmarking, quantification of phenotypes showed no benefit ahead of either of the unquantified reference sets (curated and text-mined) in ranking genes within the disease-associated genome (Figure 25). The quantification of phenotypes was expected to demonstrate the same benefit as it had in previous benchmarking experiments – no evidence was observed here but a larger sample size would be required to confirm this. The phenotype-based variant prioritisation methods were superior to variant-based prioritisation of PhenIX, consistent with phenotype-based prioritisation methods outperforming variant methods in many benchmarking investigations (Singleton et al., 2014; Smedley & Robinson, 2015; Zemojtel et al., 2014). Incorporating variant score into the phenotype score for a combined score showed no significant benefit, although this would be expected in a larger dataset.

These variant prioritisation experiments were conducted blind to the suspected mode of inheritance in these patients. Incorporating mode of inheritance into phenotype search would greatly reduce the gene search space as only relevant genes for each genotype would remain – heterozygous variation in disease genes of recessive disease would not be considered. Also, variant search space could be further reduced by filtering using pathogenicity prediction scores. Considering such scores as strong individual lines of evidence is not recommended (D. G. MacArthur et al., 2014) and therefore they were not used in the filtering schema, but in practice a variant is seldom considered if prediction tool scores unanimously

indicate non-pathogenicity. Pathogenicity prediction scores can be discordant across different tools due to the different annotations/models used, and scoring is not available for all variants, but it would be reasonable to remove variants with unanimous predictions that suggest they are benign.

Of the 55 patients without diagnostic reports who were assigned HPO terms, 7 patients were highlighted for whom the use of phenotype-based variant prioritisation identified variants of interest that warrant further investigation. In 5 of the 7 cases the variant ranked in the top 10 when sorting exome variants by gene phenotypic similarity to the patient. In each case (apart from the *NOD2* heterozygous variant) there were variants with higher gene phenotypic similarity than the candidate selected, though these were ruled out, due to either: (*i*) the variant zygosity being inconsistent with the reported mode of inheritance of phenotypes associated with the gene; (*ii*) CADD/SIFT/PolyPhen2 scores that were not suggestive of pathogenicity; (*iii*) reports of them being benign in ClinVar. As stated when discussing the benchmarking of variant prioritisation, this search space could be reduced automatically (*i*) by removing single heterozygous variants in genes causative of recessive disease. Also, (*ii*) variants with unanimous benign prediction scores could be removed, although manual appraisal of variant pathogenicity prediction scores would still be required. For example, in patient IV, both of the highlighted *FAT4* variants had CADD and PolyPhen2 scores that indicated pathogenicity, whilst SIFT predicted that they were tolerated protein changes. However, (*iii*) filtering based on ClinVar reporting would be less simple as there are often multiple reports to be considered for many variants which may conflict in their assessment of pathogenicity, and this information is not readily computationally accessed for automatic variant filtering.

It is interesting that the split between patients with and without HPO terms was relatively even between diagnosed and undiagnosed patients. Given the capabilities of HPO-based tools to either suggest candidate clinical diagnoses and prioritise genes, it is surprising that many of the undiagnosed cases were not annotated with HPO terms – in these cases, unstructured phenotypic information may have been used in individual cases to suggest clinical diagnoses or genes. Demonstrated here, the use of a HPO-based methodology of prioritising genetic variants helped identify promising candidates in 7 of the cases (and candidates in other cases not discussed) – if only a fraction of these are established as molecular diagnoses, the use of this methodology would be vindicated because it only requires the trivial effort of defining patients with HPO terms.

If there are groups of patients without machine-readable phenotypes, there are data sources available that can be leveraged to capture phenotypic data if retrospective phenotype annotation is not available. Firstly, clinic letters from patients can be text-mined to identify occurrences of HPO terms, which was demonstrated in the subset of patients (n=14) with anonymised clinic letters available. Calculating similarity to the reference set using HPO terms annotated by text mining identified the causative genes at the top rank for both individuals with a diagnostic variant. In an additional patient, a group of variants were identified within a highly phenotypically relevant gene (for an autosomal recessive disorder), though sequencing of parents is required to establish whether these occur in *trans*. HPO terms were not assigned to this patient and the identified gene was not in the virtual gene panel, so this variant may not have been highlighted if it weren't for the use of text mining. This demonstrates the value of annotating undiagnosed cases with HPO terms through automated methods in cases where manual annotation is not

possible. Additionally, the gene panels selected for each patient contain latent phenotypic information, which can be 'reverse-engineered' by identifying the OMIM phenotypes caused by each gene and extracting the HPO terms to construct a composite predicted HPO phenotype. Querying this reconstructed phenotype can help expand the gene list through consideration of similar phenotypes – this can be particularly useful when virtual gene panels were constructed years ago, and reverse-engineered phenotypes can be used to update them to include recently made discoveries.

The use of approaches such as those developed in this thesis, or tools such as PhenIX, that investigate all known human disease genes (termed the disease-associated genome) are appropriate in a diagnostic setting, which requires that identified variants are in genes with a previously described role in causing Mendelian disorders. However, in a research setting, the use of gene discovery tools such as PHIVE and hiPHIVE from the Exomiser suite (Smedley et al., 2015) may be more appropriate for the identification of candidate variants based on phenotypic similarity to both human phenotypes and animal models. This can be used in combination with phenotypic matchmakers to identify individuals across the world with similar phenotypes (and genotypes) (Buske, Girdea, et al., 2015) to provide the additional lines of evidence that variants in a particular gene are causative of the same phenotype in multiple individuals.

Chapter 5 – Utilisation of phenotype questionnaire data in a common complex disease dataset (acne) to aid interpretation of GWAS results

5.1 Introduction

Aside from rare diseases, the use of machine readable phenotypes is also becoming increasingly important in the genetic analysis of common complex disease. Several genetic studies utilise disease codes (e.g. ICD9-10-11) within electronic health records to identify disease phenotypes of interest (Dewey et al., 2016; Howard et al., 2018; Krokstad et al., 2013; Mitchell et al., 2016; The Michigan Genomics Initiative, 2016; UK Biobank, 2018) and further granularity can be achieved by using self-reported phenotypes (UK Biobank, 2018) or questionnaires filled by the attending clinician or nurse. In large scale GWAS, the necessity to identify vast numbers of cases may result in the use of an umbrella diagnosis that may cover many subphenotypes. This introduction of heterogeneity to the phenotype definition can reduce power to detect phenotype-specific effects, and analysing more homogenous populations of cases expressing subphenotypes can lead to insight into the genetic basis of such subphenotypes in GWAS analyses (Eichler et al., 2010; Kulminski et al., 2016; MacRae & Vasan, 2011). The aim of this chapter is to evaluate the potential of using phenotypic information gathered from questionnaires filled for patients with severe acne.

Acne is an inflammatory disease that affects the skin through the pilosebaceous unit – hair follicles in the skin that are associated with an oil gland. Acne primarily affects the face, chest and back, where pilosebaceous units are most densely concentrated. Acne often starts in early puberty and clinical features include

seborrhoea, lesions (inflammatory on non-inflammatory), and various degrees of scarring. Many classifications exist for acne lesions based on whether they are non-inflammatory (open and closed comedones) or inflammatory (papules, pustules, nodules and cysts) (Shalita, 2004), with nodulocystic acne representing a severe classification (Williams, Dellavalle, & Garner, 2012). The severe inflammatory response to acne can result in permanent scarring, with classifications such as ice-pick, atrophic, keloid, hypertrophic and perifollicular elastolysis (Alster & West, 1997; Jacob, Dover, & Kaminer, 2001; Varadi & Saqueton, 1970). A commonly used severity grading score is the Leeds technique, which involves counting and categorising lesions as inflammatory or non-inflammatory (Burke & Cunliffe, 1984; Purdy & de Berker, 2011).

Biological mechanisms underlying disease development are thought to involve sebum production, follicular keratinization, inflammation, and colonisation of follicles by *Propionibacterium acnes*, but are poorly understood (Williams et al., 2012). Acne impacts the psychological health of affected individuals, and is associated with depression, suicidal ideation, anxiety, psychosomatic symptoms, shame, embarrassment and social inhibition (Kubota et al., 2010), and effective treatment helps to resolve such issues (Hahm et al., 2009). Topical and systemic agents are typically used to treat severe acne, which act to suppress the microbiome repertoire or the activity of sebaceous glands, though regimes can be ineffective and cause skin irritation which may result in discontinuation (Williams et al., 2012).

Twin studies have indicated the importance of genetic factors in acne risk, with heritability estimates of 78% and 81% in Chinese and UK Caucasian populations respectively (Bataille, Snieder, MacGregor, Sasieni, & Spector, 2002; Wei et al.,

2010). GWAS have been conducted on severe acne phenotypes, identifying associations with several genomic loci, including two in Han Chinese populations (He et al., 2014) and a further 15 in European populations (Navarini et al., 2014; Petridis et al., 2018). These loci have highlighted the potential role of several genes and pathways in disease pathogenesis. Pathway analysis of genes surrounding hit loci has indicated the importance of the TGFβ pathway, which can be linked to acne pathogenesis through its involvement in keratinocyte proliferation (Navarini et al., 2014). Fine mapping of loci signals has highlighted variants likely to underlie observed association signals within *WNT10A* and *SEMA4B* (where the latter is at a critical position of a *TP63* binding motif), and regional colocalisation with skin eQTLs has identified *LAMC2* and *LGR6* as putative causal genes: all of these genes have established roles in controlling the development, morphology and activity of hair follicles (Petridis et al., 2018).

Questionnaire data was collected for patients from both Navarini *et al*. (discovery stage) and Petridis *et al*. UK GWAS studies. In this chapter, the clinical information provided is used to identify acne subphenotypes for genetic association testing, hypothesising that common genetic variation contributes to the differences between acne subphenotypes. There are several challenges in interpreting the questionnaire data, as it is not coded or represented in ontology terms, and also contains a large amount of missing data. Furthermore, any subphenotypes defined must comprise a sufficient number of individuals for statistically powerful analysis.

5.2 Materials and Methods

5.2.1 Questionnaire

Clinical information was recorded for patients from Navarini *et al*. (discovery stage) and Petridis *et al*. UK GWAS studies, for which ethical approval was obtained from the NRES Committee London-Westminster (reference CLRN 05/Q0702/114). Data collected from each patient consisted of (*i*) a case report form (CRF) filled by the attending clinician or nurse detailing diagnostic information (following assessment by a trained dermatologist) and (*ii*) self-reported information from a questionnaire filled by patients. Patients were recruited through a network of 45 dermatology centres in the UK (17 centres in the initial discovery study), which included: patients from Guy's and St Thomas' (GSTT), patients recruited from visits to other hospital sites, and patients who were sent letters by their hospital and asked to join. A different CRF form was used for non-GSTT patients (but this only differed in clinical information later marked as 'irrelevant') and there were further minor differences in the self-reporting questionnaires used for patients recruited by sending letters (rather than from hospital visits).

The full questionnaire dataset comprised a total of 162 different questions, separated into 6 broad categories:

- Patient details: Basic information such as name (anonymised to single initials), date of birth, sex and ethnicity.
- Family history: Of acne and other dermatological disorders (psoriasis, eczema, hidradenitis suppurativa).
- Diagnosis: Age and year of both diagnosis and onset.

- Acne diagnosis: Acne clinical variants such as infantile acne, nodulocystic acne, polycystic ovary syndrome acne (some of which were inclusion/exclusion criteria, listed below).
- Clinical: Anthropometric data (weight, height, BMI), a breakdown of different lesion and scarring types (and where they are situated) and disease severity scores (Leeds score and dermatology life quality index (DLQI)).
- Treatments: treatment type, date(s), dosing and response.

Among these fields were questions that comprised the inclusion and exclusion criteria, which were as follows:

Inclusion criteria (any of the following):

- A clinical diagnosis of nodulocystic acne vulgaris
- A clinical diagnosis of moderate to severe acne vulgaris requiring treatment with isotretinoin
- A clinical diagnosis of acne vulgaris with a Leeds Grading Score of more than or equal to 5 in at least one site
- A clinical diagnosis of either acne conglobata, acne fulminans, sandpaper acne OR submarine acne (clinical subtypes of acne vulgaris)

Exclusion criteria:

- Be unable to give informed consent
- Blood transfusion received within last 4 weeks
- Any evidence of acne agminata, acne rosacea, or other acne form eruption not fulfilling the inclusion / exclusion criteria
- Acne associated with evidence of virilisation or other significant hormonal abnormality based on clinical assessment by a trained dermatologist

- Drug induced acne (determined by a trained dermatologist) e.g. steroid induced acne
- Patients with acne due to their occupation
- Patients with acne due to working with halogenated hydrocarbons
- Body builders

5.2.2 Data filtering

To include only data that is phenotypically relevant, and that can also be simply converted to machine-readable format (e.g. not free-text), the following filtering steps were applied to the dataset:

- Removal of questionnaires in instances where the same individual had two or more completed questionnaires completed at different times added to the databases. All such questionnaires were removed to avoid issues of possible discrepancy between different responses (n=106).
- Removal of questionnaires that did not originate from individuals contained in the final analysis of the initial Navarini *et al*. study or the subsequent Petridis *et al*. study (from here named DS1 and DS2 respectively).
- Removal of questions not containing useful information pertaining to acne clinical presentation. These were in one of three broad categories (removed questions and reasons listed in Appendix 2):
    - Basic patient details with no phenotypic relevance (e.g. “name”, “date of visit”)
    - Questions with free-text responses which cannot be computationally processed and incorporated into similarity models (e.g. “acne other specify”, “lesion overall”, “scarring overall”).

    - Questions detailing phenotypic information not related to the acne phenotype (e.g. “eczema family history”, “psoriasis family history”).
- Removal of questions with monomorphic responses. This required initially standardising uninformative responses to single value (e.g. “unknown”, “incomplete”, “indeterminate”, “[BLANK]”) before assessing whether responses were monomorphic for a question. Binary questions were considered to have monomorphic responses if they only contained “No” and uninformative responses, which was considered appropriate due to the nature of questionnaires, which asked for *presence* of acne clinical variants or lesions/scarring types. This filtering step resulted in the removal of all study exclusion criteria (as only “No”/blank responses remained after selecting individuals used in final analysis of the genetic studies).
- Removal of poorly filled questions. Questions where <1% of individuals responded were removed. 1% is a somewhat arbitrary threshold and only resulted in the removal of biometric data (height, weight, BMI) which was available for only 0.6% of genotyped individuals.

Questions included after filtering are listed in Appendix 3.

5.2.3 Missingness clustering

Responses were clustered by missingness to assess structure of the data that was due to different patterns of questionnaire missingness rather than phenotype similarity. To perform this data was recoded to binary format based on the presence or absence of response.

5.2.3.1 By question

Question vectors were clustered by missingness using hierarchical clustering, which considers all vectors as separate clusters, and in each iteration of the algorithm the two closest clusters are combined until there is only one cluster remaining. A Euclidean distance measure between question vectors was used and the distance between clusters was defined as the minimum between any of their constituent vectors. Questions were visualised on a heatmap coloured by Pearson correlation coefficients.

5.2.3.2 By patient

Principle component analysis (PCA) was used to identify clusters of individuals for which the same parts of the questionnaire had been filled. PCA transforms a dataset of possibly correlated variables into a smaller number of uncorrelated variables (principle components). Each sequential principle component explains as much of the variance as possible, enabling lower dimension representations of datasets with high dimensionality.

5.2.4 Response clustering

Three data types exist in the questionnaire dataset: binary, quantitative and factors (categorical variables). Binary response data was simply encoded as 0 (“No”) and 1 (“Yes”). Quantitative data was linearly normalised to between 0 and 1. Factor data was one-hot encoded (i.e. the creation of separate vectors for each unique response value, each with binary encoding).

Missing values are not permitted by clustering algorithms, and this can be resolved by either removing cases with missing values or imputing missing values (or a combination of the two). To retain the maximum number of individuals with

clinical information the following imputation strategy was employed for the different data types:

- Missing binary data was encoded as 0 – assuming missing data is equivalent to a "no" response.
- Missing quantitative data was filled with the median value.
- There were no missing values for factor variables.

5.2.4.1 By question

After recoding the response data, questions were again clustered using hierarchical clustering and visualised on a heatmap coloured by Pearson correlation coefficients.

5.2.4.2 By patient

5.2.4.2.1 t-SNE

To identify patient clusters for genetic study, PCA was not sufficient to reduce the data to a two-dimensional representation as it explained only 37% of the dataset variance. Therefore, t-distributed stochastic neighbour embedding (t-SNE) was employed – t-SNE enables high-dimensional datasets to be visualised by computing low-dimensional embeddings where pairwise distances between data points are reflective of pairwise distances between the high-dimensional input. (Van Der Maaten & Hinton, 2008). t-SNE was performed on the first 16 principle components of the original data, the minimum number of principle components to explain 85% of the variance.

5.2.4.2.2 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) was used to cluster the data based on the two-dimensional t-SNE embeddings. DBSCAN is a

widely used density-based clustering algorithm which uses the following parameters: *minimum points* – the number of data points within a neighbourhood required to define a cluster (here: 100); ε – the maximum distance between two points to be considered neighbours (here: 8).

5.2.5 Power calculation

For each explicitly defined binary subphenotype within the questionnaire data, genome-wide association testing could be conducted using the following cohorts as controls:

- Unselected population controls from original Navarini *et al*. and Petridis *et al*. studies.
- Individuals within the case cohort for whom the subphenotype was explicitly stated not to be present (i.e. "No" responses).
- All individuals within the case cohort who weren't stated to have the subphenotype (i.e. all non-"Yes" responses).

For each potential genome-wide association test (subphenotype vs. each of the respective control populations) statistical power was calculated using equations implemented in the GAS power calculator (Skol, Scott, Abecasis, & Boehnke, 2006). Using the number of cases and controls for each potential analysis, the power to detect a SNP with genome-wide level of significance of association ($P = 5 \times 10^{-8}$) was calculated, looking for a SNP with 50% allele frequency and genotype relative risk of 2 (assuming a multiplicative model and a disease prevalence of 0.1%). Analyses with less than 80% power were not conducted.

5.2.6 GWAS

Genome-wide association testing was performed for each case/control group with sufficient statistical power (as defined above) using the same meta-analysis pipeline followed in Petridis *et al*.

Genotyping, quality control and imputation (summarised from Navarini et al. and Petridis et al.)

In DS1, genome-wide genotyping was conducted on Illumina HumanOmniExpress-12v1_H microarrays for cases and on Illumina 1.2M platform (at the Sanger Institute, Cambridge) for controls, which were either healthy blood donors from the United Kingdom Blood Service (UKBS, n=2,478) or individuals from the 1958 Birth Cohort (58C, n=2,661). Genotyping was conducted in two batches in DS2 cases, using the Human 250 Illumina Omni Express Exome 8v1-2 (2,567 cases) and Illumina Omni Express Exome 8v1-3 (1,961 cases) and control genotypes were obtained from the English Longitudinal Study of Aging (ELSA, genotyped on the Illumina Human Omni 2.5) and the Understanding Society Project (USP, genotyped on the Illumina Human Core Exome v12.0). All control cohorts used were unselected population control cohorts. Quality control was performed in the respective datasets using accepted quality control procedures (Weale, 2010), considering call rate, Hardy-Weinberg equilibrium, discrepancy of inferred gender with sample information, excessive heterozygosity, cryptic relatedness and ancestry outliers (Navarini et al., 2014; Petridis et al., 2018). QC was performed in two batches within DS2: one containing genotypes from 2,567 cases and 7,452 controls from ELSA and the other with 1,961 cases and 9,500 controls from USP. Phasing and imputation was undertaken using the Haplotype Reference Consortium (HRC version r1.1) reference panel on

the Michigan Imputation Server (Das et al., 2016). After imputation, SNPs with R^2 < 0.7 were excluded. In this testing SNPs with a minor allele frequency of < 5% in either study batch were excluded from downstream analysis.

Association

Association testing was performed within DS1 and DS2 with a logistic Wald association test (EPACTS), including the first four principal components (as well as the QC/imputation batch in DS2 only) as covariates. For subphenotype analyses with insufficient numbers of cases or controls (<10 of either) in DS1 or DS2 for stable association testing on EPACTS, summary statistics were reported for analysis in the single dataset, and meta-analysis was not conducted. This consideration was reflected in power calculations.

Meta-analysis

Results from the DS1 and DS2 association analyses were incorporated into a standard error-weighted meta-analysis with GWAS summary statistics performed with METAL (release 2011-03-25) (Willer, Li, & Abecasis, 2010).

5.3 Results

5.3.1 Data filtering

The original acne patient questionnaire dataset contained 7,911 patients and responses for up to 162 questions. 5,476 patients remained after removing any duplicate records and selecting only individuals that were included in the final analysis of Navarini *et al*. (DS1) and Petridis *et al*. (DS2). 1,747 of 1,779 of DS1 individuals and 3,729 of 3,823 DS2 individuals (98% of entire dataset) had questionnaire data without duplicate entries. After all filtering 66 questions remained (Table 24).

Table 24: Acne questionnaire dataset size following each filtering step. The number of questions removed for each reason may not be the correct number of questions with that property (i.e. some questions defined as irrelevant were also monomorphic or poorly filled and were therefore removed before that step).

Stage	*Questions*	*Patients*
Original dataset	162	7955
Remove duplicate entries	162	7849
Genotyped patients only	162	5476
Remove irrelevant questions	81	5476
Remove monomorphic response questions	69	5476
Remove poorly filled questions (<1%)	66	5476

5.3.2 Response Rates

The response rates differed greatly between DS1 and DS2, with questions filled at a much higher rate in DS2 individuals on average (Figure 33). “Sex” and “ethnicity” data fields were fully populated in both datasets, while “hirsutism” and “androgenic alopecia” clinical variants were largely left blank across datasets. “Age of onset”, “nodulocystic acne”, “acne before puberty” and the Leeds score questions were filled with reasonable rates that were comparable across datasets, and the remaining questions were filled much more in DS2 than DS1 (Figure 34) – this group mostly consisted of the lesion and scarring information.

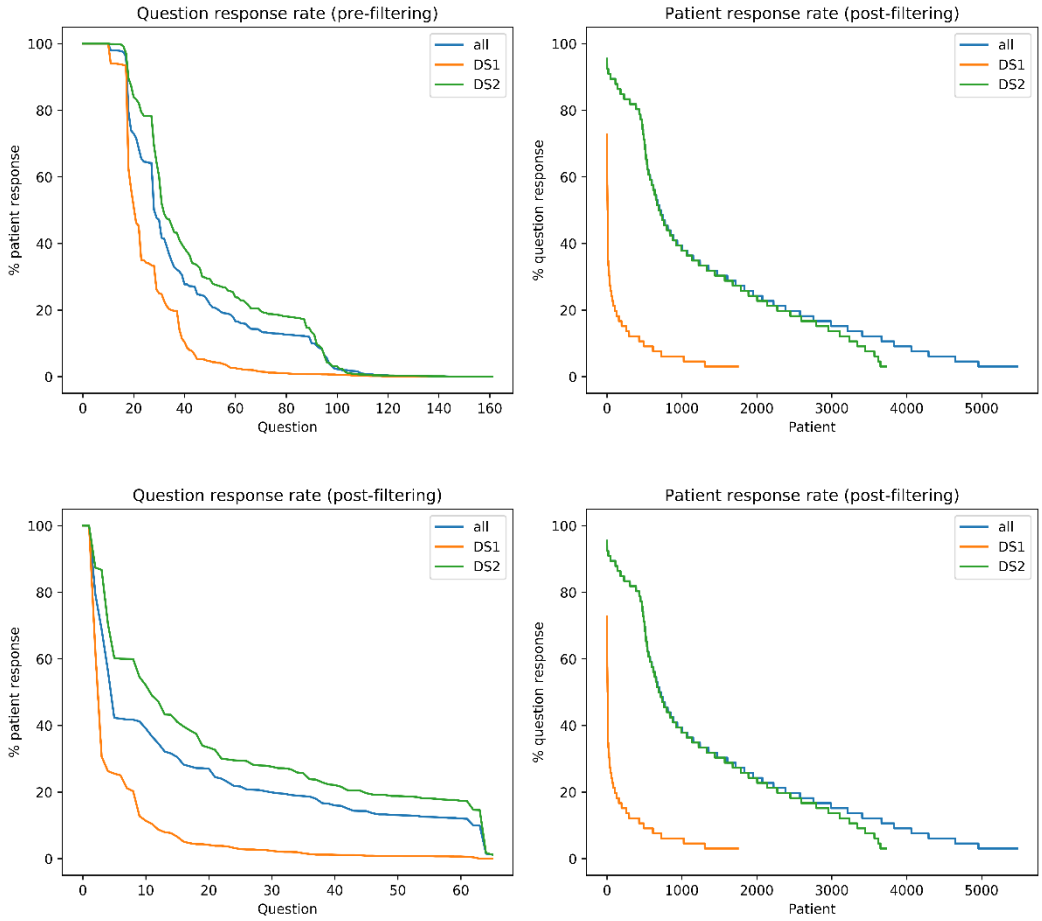


Figure 33: Questionnaire response rates by question and patient (left and right respectively), before and after data filtering (top and bottom respectively).

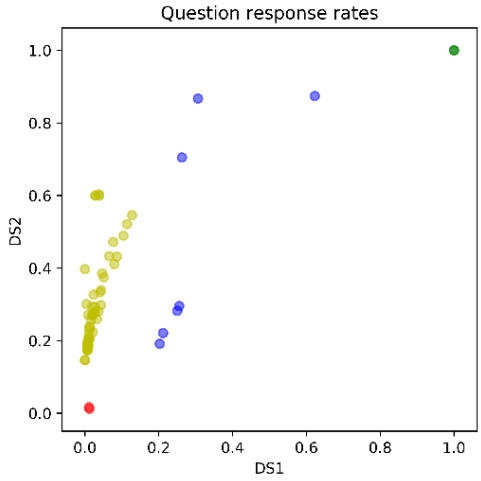


Figure 34: Comparison of question response rates (after data filtering) between dataset 1 and 2 (n=33). Green: questions filled universally highly; Red: questions with low response rates in both datasets; Blue: questions with comparable response rates; Yellow: questions filled to a much greater degree in DS2 than DS1.

5.3.3 Missingness clustering

5.3.3.1 By patient

When plotting the first two principal components of the data (explaining 53.1% of the variance; Figure 35) two populations emerged – referred to as population 1 and 2 (green and grey respectively in Figure 36). Population 1 (n=4,839) contained individuals from both genotyped datasets, while population 2 (n=637) almost exclusively comprised individuals from dataset 2. There was a dramatic difference in mean overall questionnaire fill rates between these populations (18.0% and 78.6% respectively), which was most evident in scarring data (as well as lesion data to a lesser degree). This difference largely corresponded to the presence of more "No" values (Appendix 4).

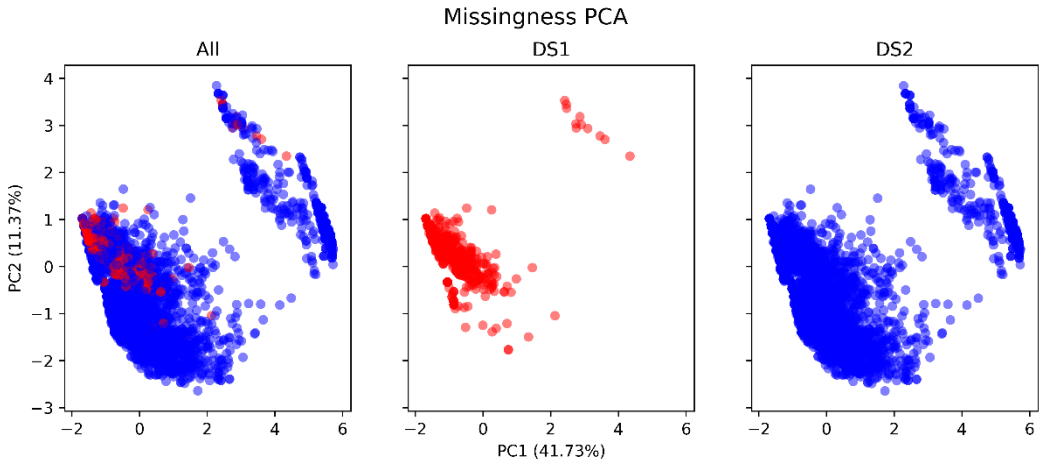


Figure 35: PCA plot for missingness of acne patient questionnaire responses. DS1 patients are shown alone in the central panel, DS2 patients are shown alone in the right panel, and both groups of individuals are shown together in the left panel (DS1: red; DS2: blue).

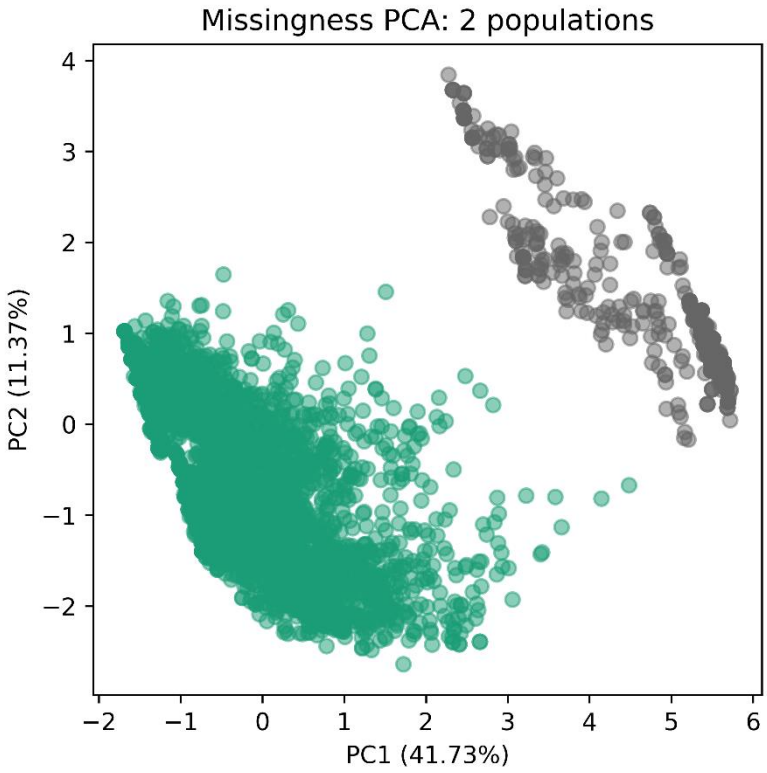


Figure 36: PCA plot for missingness of acne patient questionnaire responses (equivalent to left panel of Figure 35), showing the two patient populations that emerged based on patterns of missingness (coloured green and grey).

5.3.3.2 By question

Hierarchical clustering and Pearson correlation coefficients were used to visualise patterns of missingness among different groups of questions (Figure 37). Spatial lesion and scarring data was plot separately (Figure 38, Figure 39).

There were certain "blocks" of questions that tended to be completed in similar groups of patients, which included lesion data (summary), scarring data (summary), Leeds scores (spatial and overall) and clinical variants within the inclusion criteria (nodulocystic, sandpaper, submarine, fulminans, conglobata) (Figure 37). Age of onset response was correlated with indication of onset before puberty. Hirsutism and androgenic alopecia had no correlation with the response

patterns of many of the other questions as they represented the minor difference in the self-reporting questionnaire for the patients recruited by letter correspondence. Note that correlation between response patterns does not necessarily indicate that responses themselves were similar.

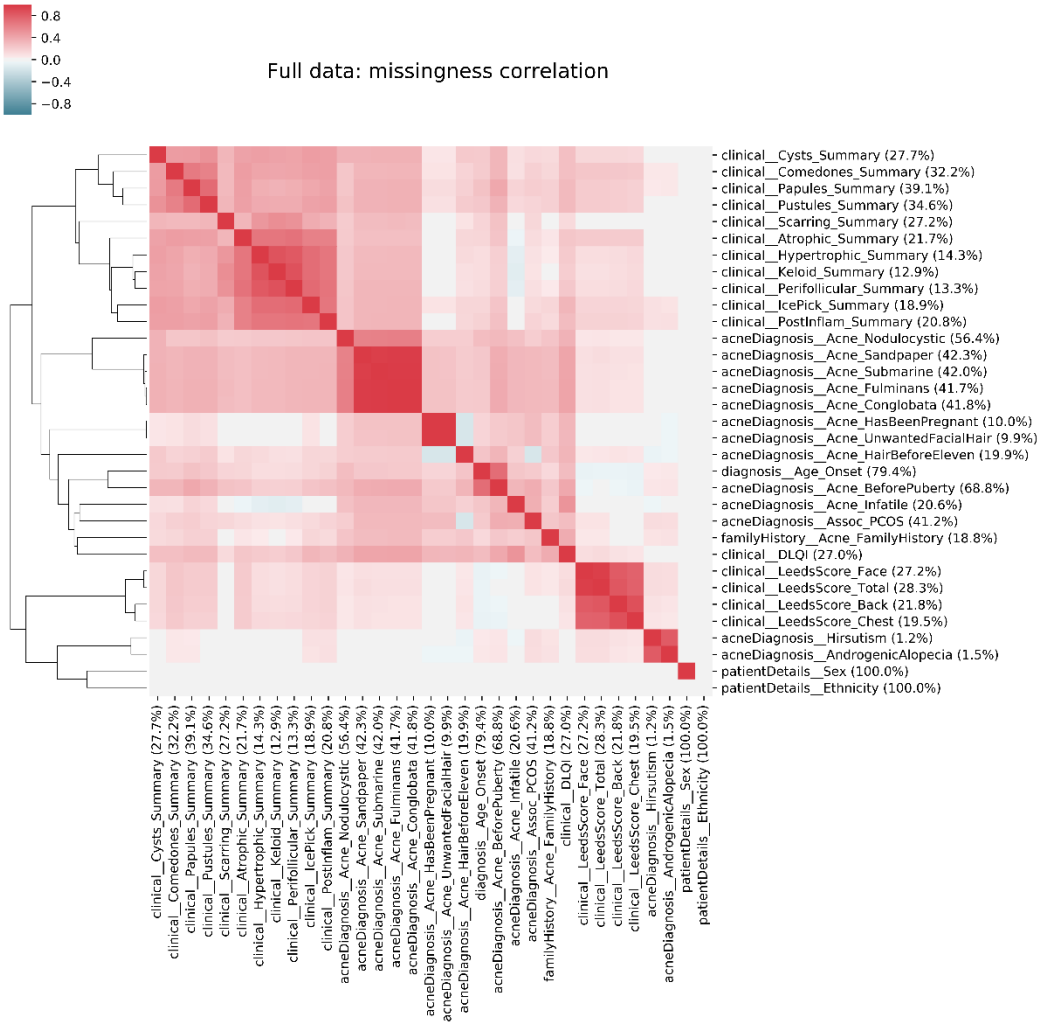


Figure 37: Correlation (Pearson) between question response missingness (ordered after hierarchical clustering). Spatial lesion and scarring data was plot separately. % value for each question corresponds to the total response rate.

With the exception of cysts, lesion types were filled more consistently across spatial locations than lesion type, though correlations were still observed across lesion types (Figure 38).

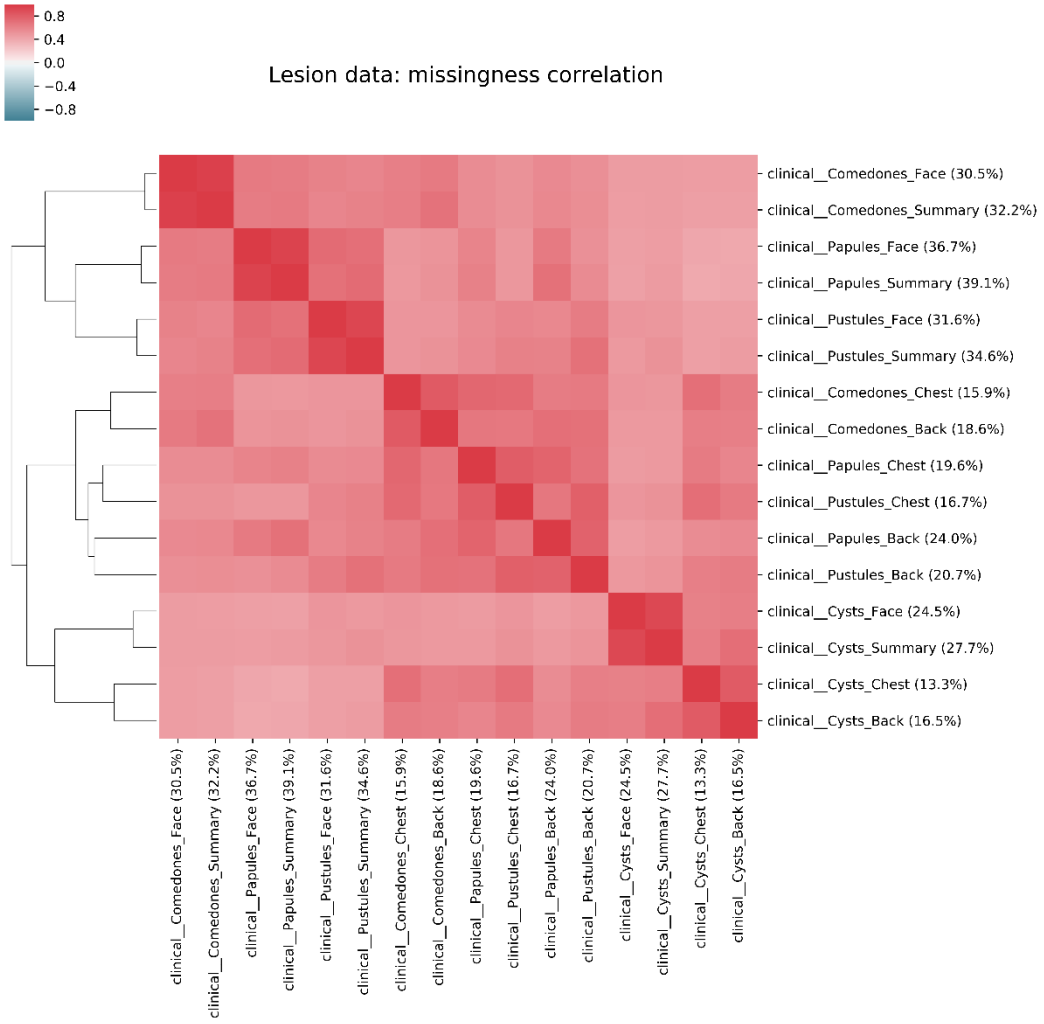


Figure 38: Correlation (Pearson) between lesion data missingness (ordered after hierarchical clustering). % value for each question corresponds to the total response rate.

The response patterns across many scarring types and locations was highly correlated, with the exception of "face" and "summary" data for atrophic, ice-pick, postinflammatory and generic scarring questions (Figure 39).

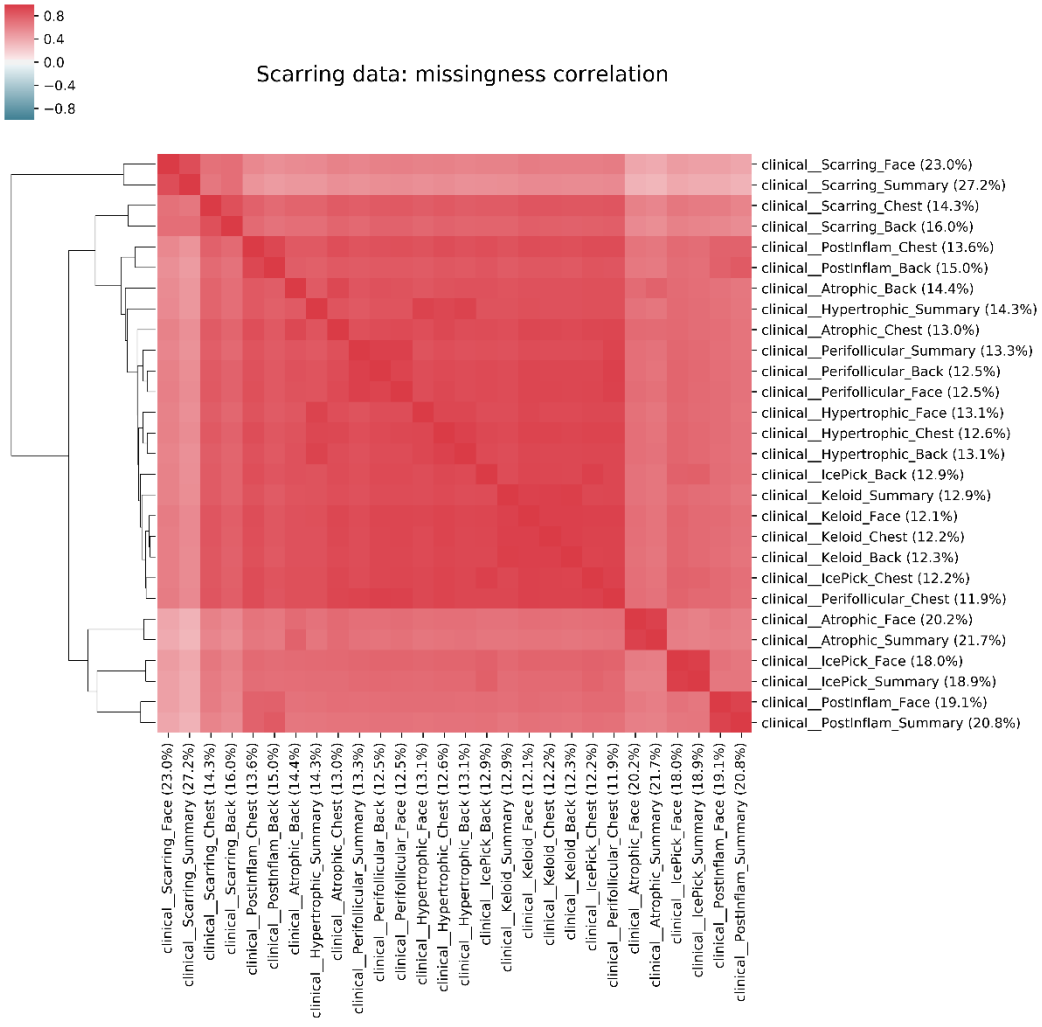


Figure 39: Correlation (Pearson) between scarring data missingness (ordered after hierarchical clustering). % value for each question corresponds to the total response rate.

5.3.4 Response clustering

After encoding data appropriately and imputing missing values according to 5.2.4, clustering was performed according to specific response rather than presence or absence of data.

5.3.4.1 By question

Structures of missingness were still apparent for lesion and scarring types after encoding responses, although structure between inclusion criteria no longer existed as only one of these was required for inclusion in the study (Figure 40). Predictably, age of onset became inversely correlated with indication of acne before puberty after encoding the data. Presence of cysts and indication of nodulocystic acne were correlated with each other, and were correlated with male acne, as were higher Leeds scores and “hair before eleven”. Higher DLQI scores, “has been pregnant”, “unwanted facial hair” and “acne before puberty” were correlated with female acne.

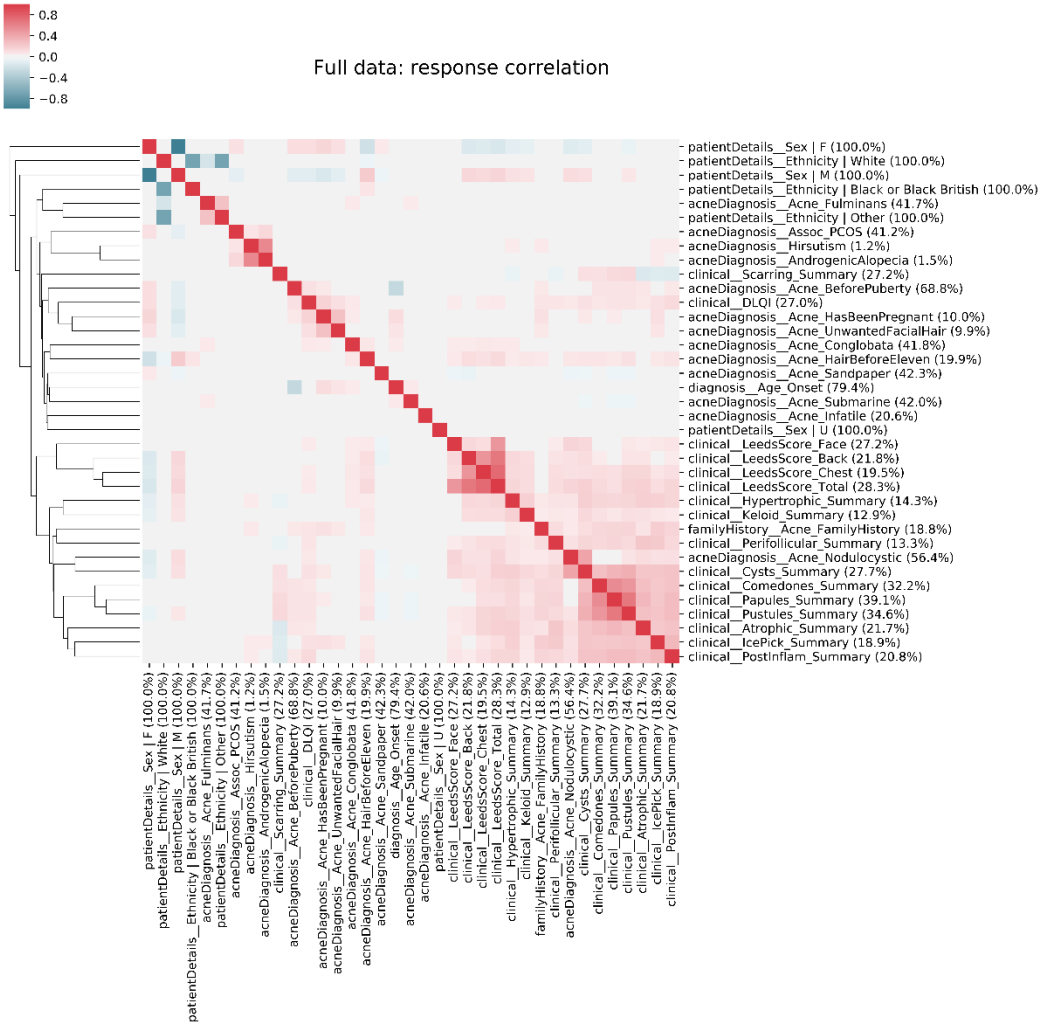


Figure 40: Correlation (Pearson) between question response values (ordered after hierarchical clustering). Spatial lesion and scarring data was plot separately. % value for each question corresponds to the total response rate.

Missingness structure remained after encoding responses for spatial and summary lesion data (Figure 41).

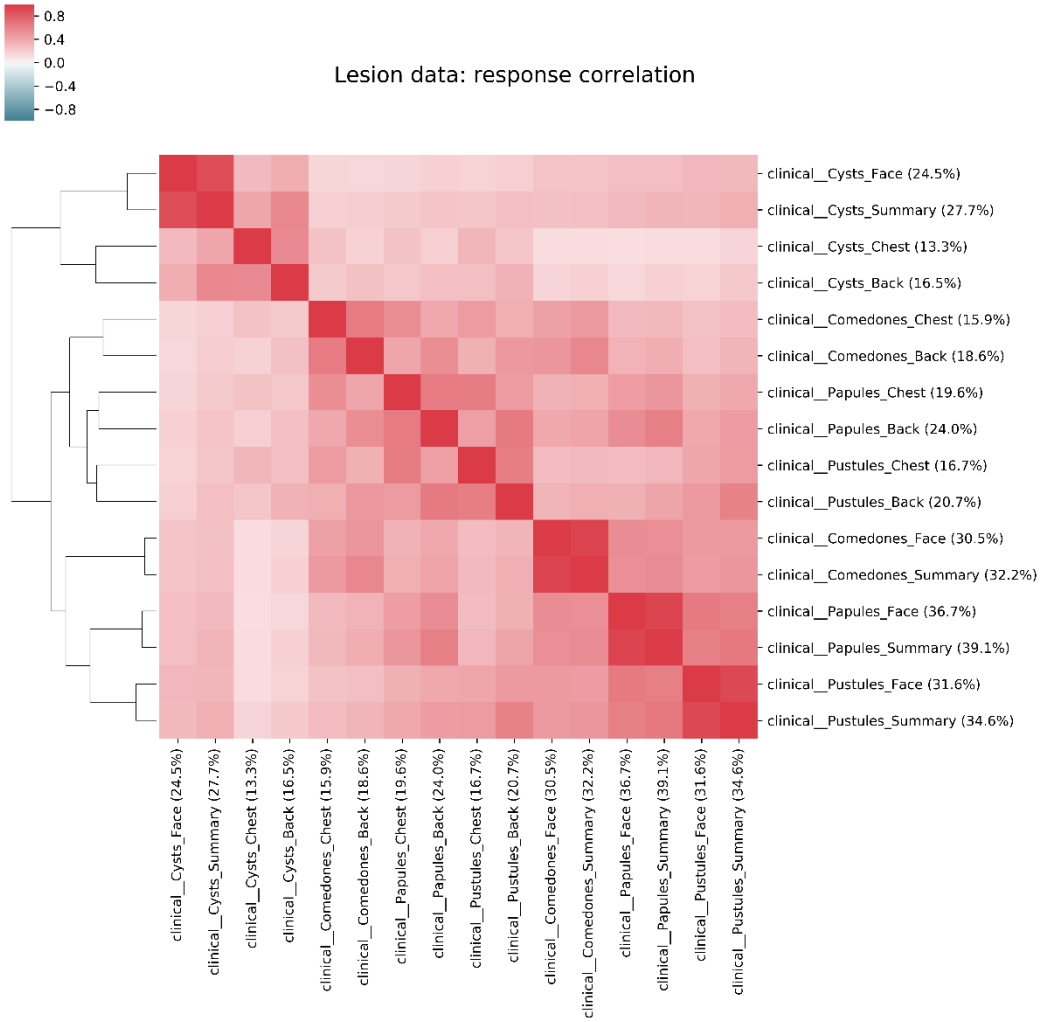


Figure 41: Correlation (Pearson) between lesion response values (ordered after hierarchical clustering). % value for each question corresponds to the total response rate.

After encoding scarring data much of the missingness structure disappeared (Figure 42), as there were many “No” responses for these questions – particularly in the well-filled population (Appendix 4). Responses were correlated more within subtypes rather than the location affected.

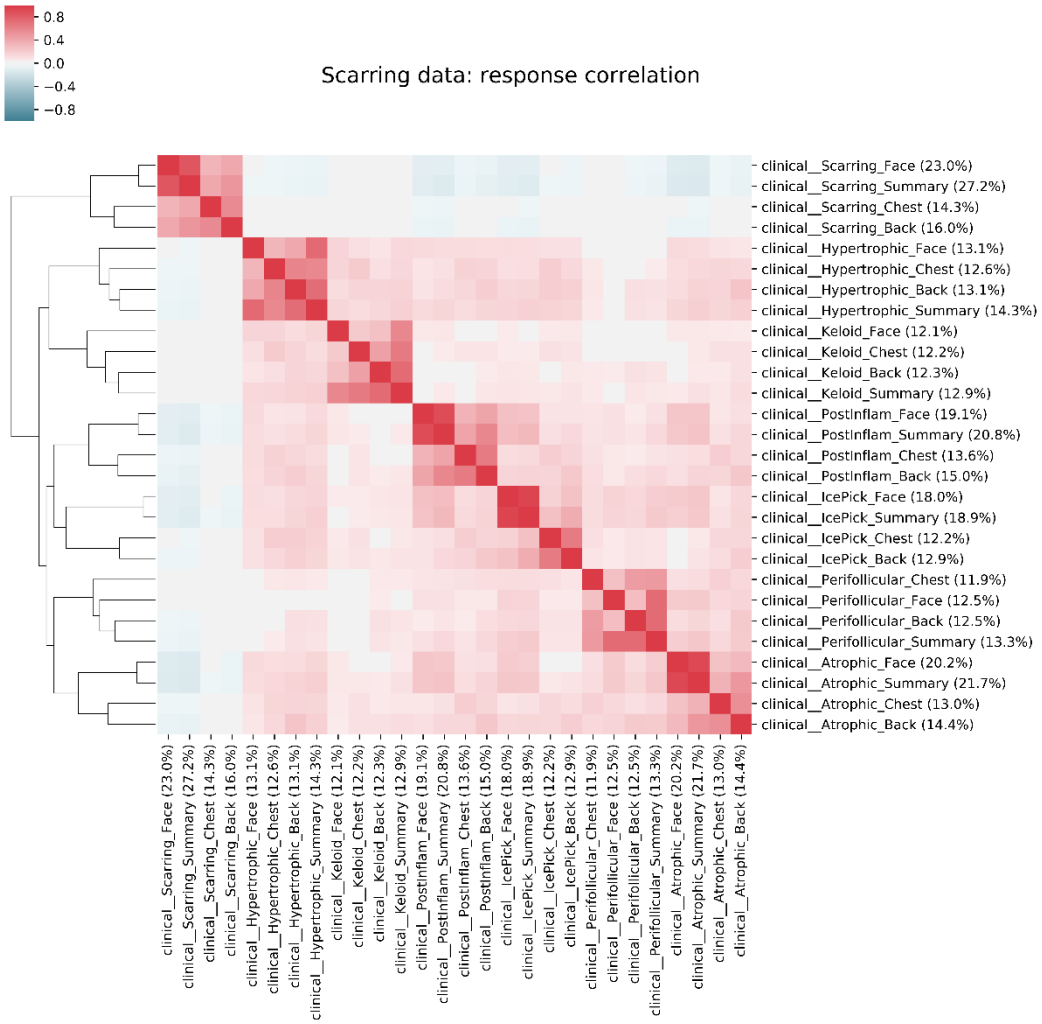


Figure 42: Correlation (Pearson) between scarring response values (ordered after hierarchical clustering). % value for each question corresponds to the total response rate.

5.3.4.2 By patient

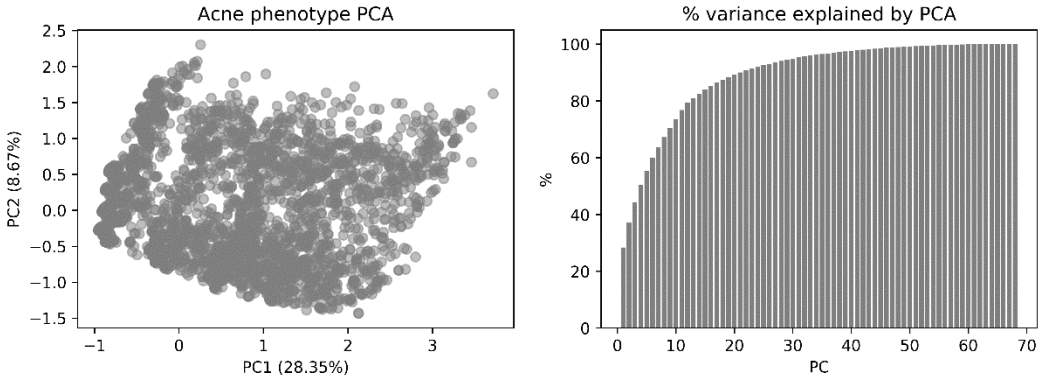


Figure 43: PCA plot of acne questionnaire data after response encoding and imputation (left) and % cumulative variance shown by each principle component (right).

PCA dimensionality reduction did not reveal any obvious clusters and explained only 37% of the variance (Figure 43). Therefore, dimensionality reduction was performed using t-SNE using 16 principal components as input (which explained 85% of the variance: additional principle components contribute minimal additions to the explained variance). Clusters were identified in the two-dimensional embeddings using DBSCAN (Figure 44). 12 clusters in total were identified with the parameters used. Four of these clusters were of interest, with clusters 1 and 3 containing a large number of densely concentrated individuals, and clusters 2 and 4 containing large numbers of patients less densely packed but distinct from the rest of the dataset. However, when scrutinising these clusters, they only corresponded to individuals with largest amounts of missing data. Of the densely packed clusters, cluster 1 (blue) corresponds to female individuals with largely empty questionnaires, and cluster 3 (purple) corresponds to male individuals with largely empty questionnaires. Of the distal clusters, cluster 2 (green) corresponds to male individuals with empty questionnaires but indication of nodulocystic acne (the most common clinical variant of acne contained in the study, Appendix 5), and

cluster 4 (red) corresponds to male individuals with empty questionnaires but indication of nodulocystic acne.

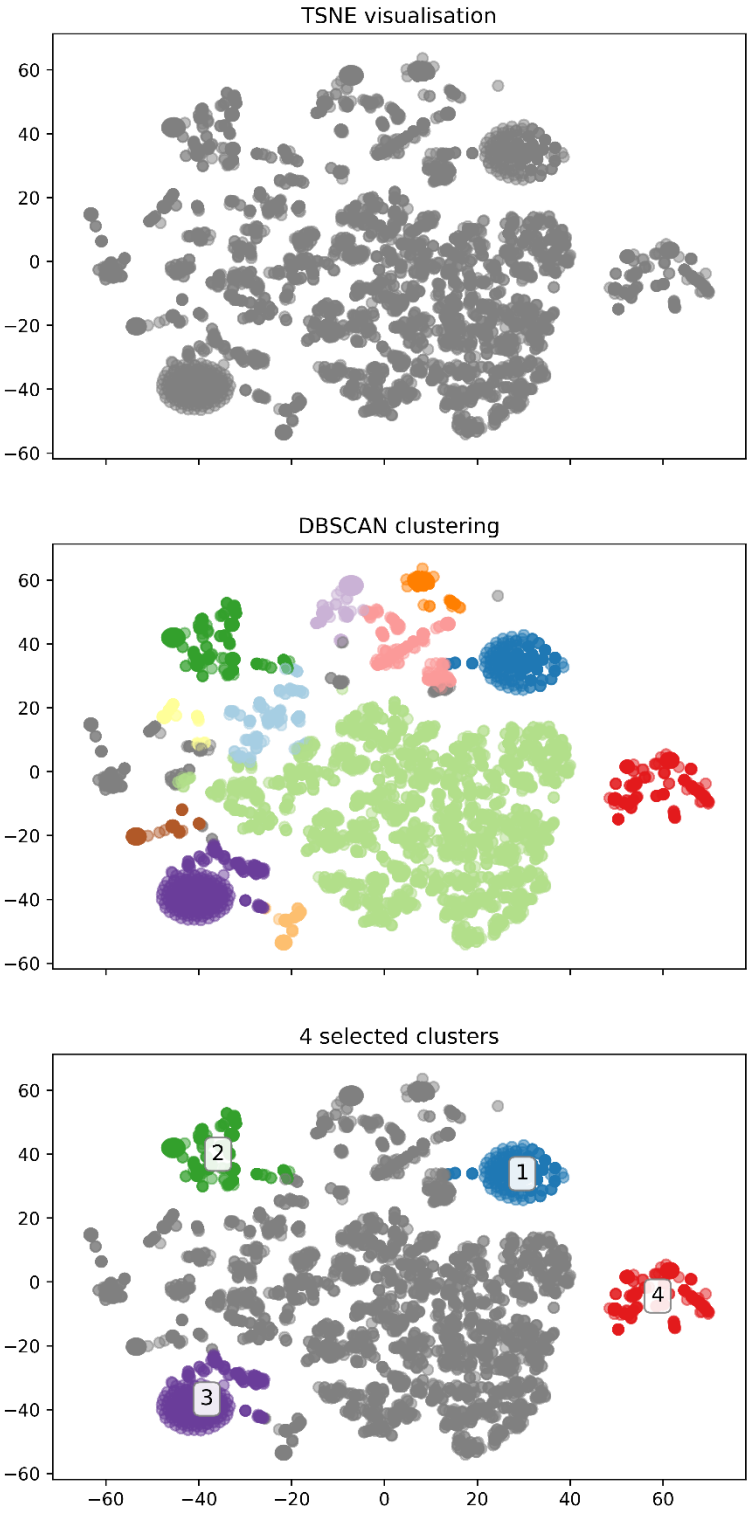


Figure 44: t-SNE plot of acne questionnaire data after response encoding and imputation. Top: two-dimensional embeddings after running t-SNE. Middle: 12 patient clusters (indicated by colour) identified using DBSCAN on the two-dimensional embeddings of the top plot (grey: individuals not included in a cluster). Bottom: 4 interesting clusters from the middle plot remain coloured, while the remaining individuals are coloured grey.

5.3.5 GWAS on binary variables

Due to the challenges of robustly identifying putative subphenotypes across all responses using dimensionality reduction and density-based clustering techniques, subsetting of the cohort was undertaken using binary responses to individual questions. Cases were defined as individuals with "Yes" responses to the question and these were compared against controls, which can be defined in a number of ways: (*i*) controls from original tested datasets, (*ii*) individuals where the trait was explicitly stated as absent (i.e. "No" responses) and (*iii*) all individuals for whom the trait was not stated as present (i.e. all non-"Yes" responses). Only analyses where there was sufficient power to detect genome-wide association of SNPs were undertaken (32 analyses in total; Appendix 5). To restrict the number of analyses conducted, spatial lesion and scarring data were removed and only the summary questions for these data were considered. 14 genome-wide association signals were identified across the 6 subphenotypes, within 6 of the known acne susceptibility loci and a novel locus not previously associated with acne (Table 25).

Table 25: Genome-wide significant hits for binary subphenotype analysis. Controls column states which cohort was used as controls: C – controls from original studies; N – "No" responses for question; N+X – all non-"Yes" responses for the question (for which no GW hits were found). Previously discovered loci from Petridis *et al*. were indicated with * in the band column.

Subphenotype	*Controls*	*Band*	*Lead SNP*	*Freq*	*P-val*	*OR (95% CI)*
familyHistory: Acne_FamilyHistory	C	4q27-28.1*	rs216101	0.2967	4.96E-08	0.75 (0.65-0.86)
familyHistory: Acne_FamilyHistory	C	15q26.1*	rs908045	0.5478	5.79E-09	1.38 (1.27-1.48)
acneDiagnosis: Acne_Nodulocystic	C	2p16.1*	rs7560605	0.5844	2.02E-09	1.23 (1.17-1.30)
acneDiagnosis: Acne_Nodulocystic	C	5q11.2*	rs626726	0.3326	3.85E-08	0.82 (0.75-0.89)
acneDiagnosis: Acne_Nodulocystic	C	11p15.3-15.2*	rs4757109	0.887	3.59E-08	0.75 (0.65-0.85)
acneDiagnosis: Acne_Nodulocystic	C	15q26.1*	rs34560261	0.833	1.73E-08	1.40 (1.28-1.51)
clinical: Comedones_Summary	C	1q25.3*	rs6703054	0.4927	9.83E-09	0.80 (0.73-0.88)
clinical: Papules_Summary	C	15q26.1*	rs34560261	0.8338	4.40E-08	1.37 (1.25-1.48)
clinical: Pustules_Summary	C	1q25.3*	rs640069	0.4938	3.40E-08	0.82 (0.75-0.89)
clinical: Pustules_Summary	C	5q11.2*	rs626726	0.3317	3.93E-08	0.81 (0.74-0.89)
clinical: Pustules_Summary	C	11p15.3-15.2*	rs61877990	0.6777	8.14E-10	0.79 (0.72-0.87)
clinical: Pustules_Summary	C	15q26.1*	rs908045	0.5471	1.27E-08	1.26 (1.18-1.34)
clinical: Cysts_Summary	C	5q11.2*	rs626726	0.3311	2.44E-08	0.78 (0.70-0.87)
clinical: Cysts_Summary	N	10q26.13	rs74158472	0.0763	1.34E-08	0.40 (0.08-0.72)

5.3.6 Hits within previously discovered loci

Restricting the cases of the original GWAS meta-analysis to only individuals with family history of acne identified two of the 15 previously discovered loci as genome-wide significant, and both of these signals appear to have significantly larger odds ratios compared to the original analysis (Figure 45). These signals are not replicated when comparing against inviduals within the acne cohort for which there was no family history (either explicitly stated or comparing against all remaining individuals).

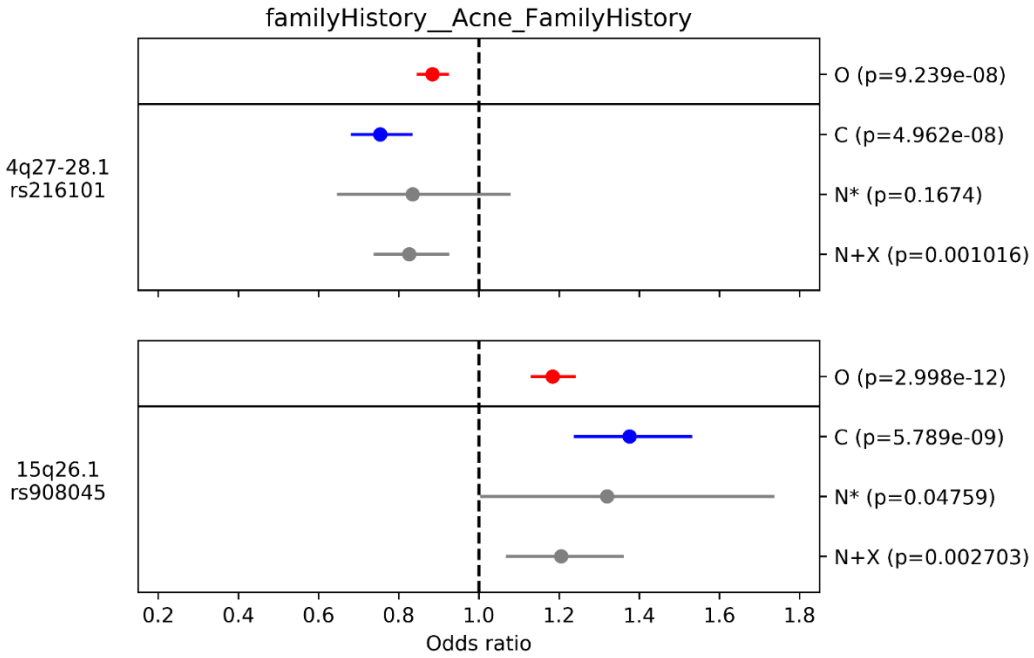


Figure 45: Loci where genome-wide significant levels of association were identified in individuals with family history of acne. Odds ratios are shown for the SNP in both the original study meta-analysis (O; red) and each subphenotype analysis undertaken here (where genome-wide significant hits are shown in blue): C – using controls from original study; N – using "No" responses as controls; N+X – using non-"Yes" responses as controls. If there were insufficient numbers of individuals to conduct one arm of the meta-analyses, the summary statistics were reported for the single arm where numbers were sufficient (*).

Restricting the original meta-analysis to only cases of nodulocystic acne identified four of the previously discovered loci at genome-wide significance, but the odds ratios were not significantly greater at any locus (Figure 46). These signals were

not replicated when comparing against individuals without nodulcystic acne, whether stated explicity or including individuals with no response.

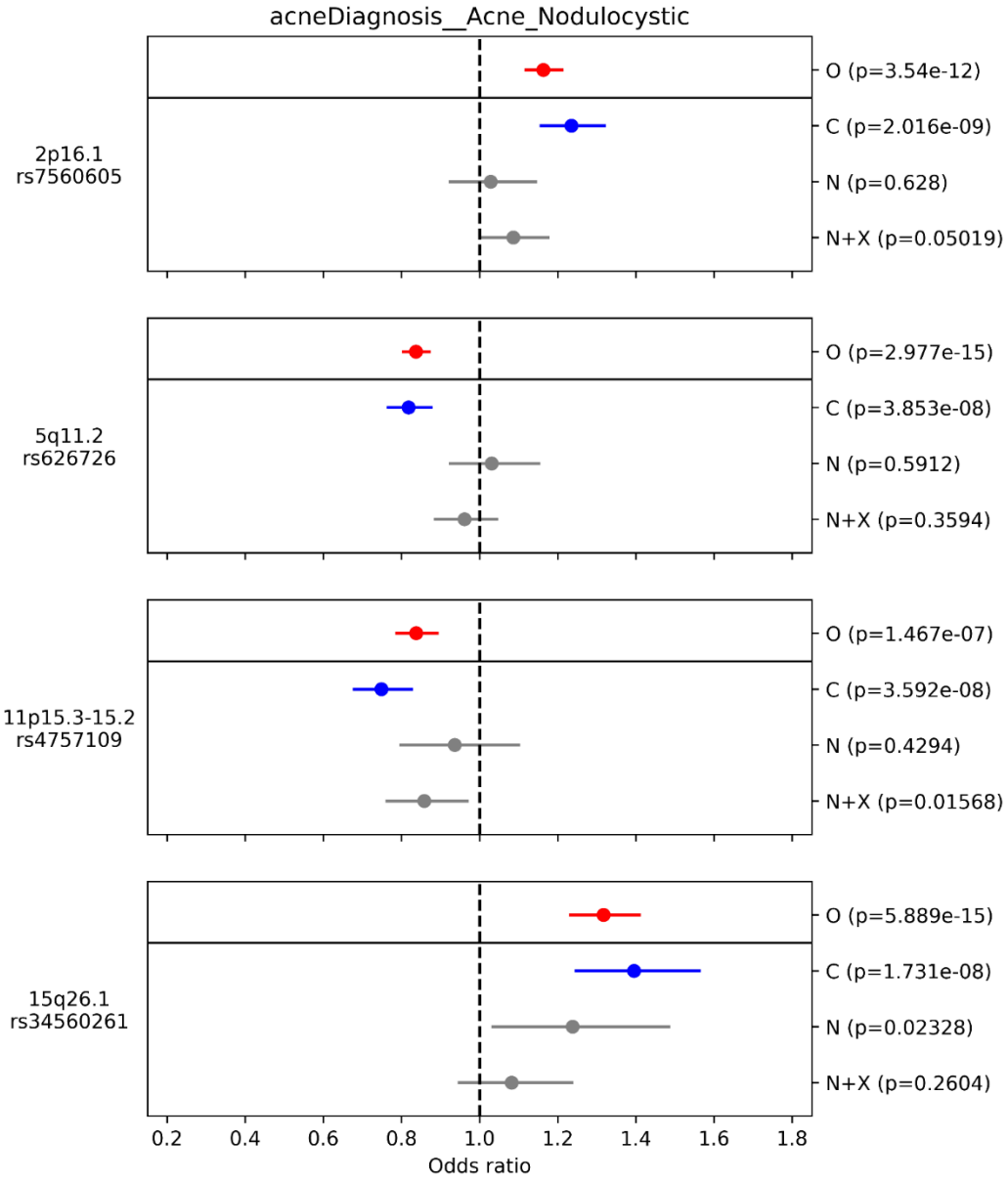


Figure 46: Loci where genome-wide significant levels of association were identified in individuals with nodulocystic acne. Odds ratios are shown for the SNP in both the original study meta-analysis (O; red) and each subphenotype analysis undertaken here (where genome-wide significant hits are shown in blue): C – using controls from original study; N – using "No" responses as controls; N+X – using non-"Yes" responses as controls. If there were insufficient numbers of individuals to conduct one arm of the meta-analyses, the summary statistics were reported for the single arm where numbers were sufficient (*).

Restricting the original meta-analysis to only individuals with comedone lesions identified one of the previously discovered loci at genome-wide significance, but the odds ratio was not significantly greater than that of the original hit (Figure 47). Again, this signal did not replicate when comparing against individuals without comedones, whether stated explicity or including individuals with no response.

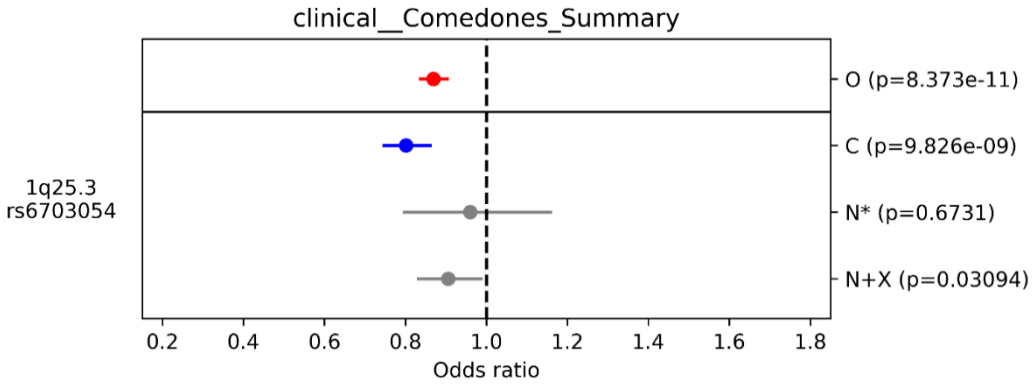


Figure 47: Locus where genome-wide significant levels of association were identified in individuals with comedone lesions. Odds ratios are shown for the SNP in both the original study meta-analysis (O; red) and each subphenotype analysis undertaken here (where genome-wide significant hits are shown in blue): C – using controls from original study; N – using "No" responses as controls; N+X – using non-"Yes" responses as controls. If there were insufficient numbers of individuals to conduct one arm of the meta-analyses, the summary statistics were reported for the single arm where numbers were sufficient (*).

Restricting the original meta-analysis to only individuals with papule lesions identified one of the previously discovered loci at genome-wide significance, but the odds ratio was not significantly greater than that of the original hit (Figure 48). This signal did not replicate when comparing to individuals without indication of papules (i.e. including missing data), and interestingly, an opposite direction of effect is seen when comparing to individuals stated to not have papules, though this is not statistically significant.

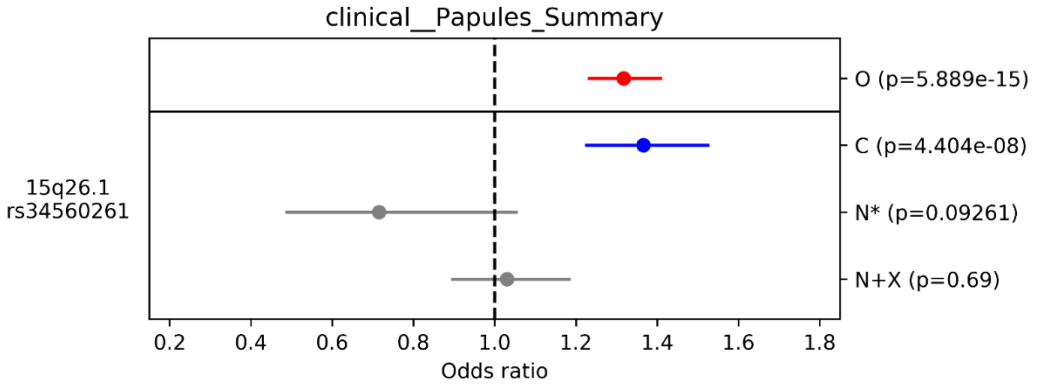


Figure 48: Locus where genome-wide significant levels of association were identified in individuals with papule lesions. Odds ratios are shown for the SNP in both the original study meta-analysis (O; red) and each subphenotype analysis undertaken here (where genome-wide significant hits are shown in blue): C – using controls from original study; N – using "No" responses as controls; N+X – using non-"Yes" responses as controls. If there were insufficient numbers of individuals to conduct one arm of the meta-analyses, the summary statistics were reported for the single arm where numbers were sufficient (*).

Restricting the original meta-analysis to individuals with pustule lesions identified four of the previously discovered loci at genome-wide significance, but the odds ratios were not significantly for any locus (Figure 49). Again, these signals were not replicated when comparing to individuals who were identified as not having pustule lesions, whether stated explicity or including individuals with no response.

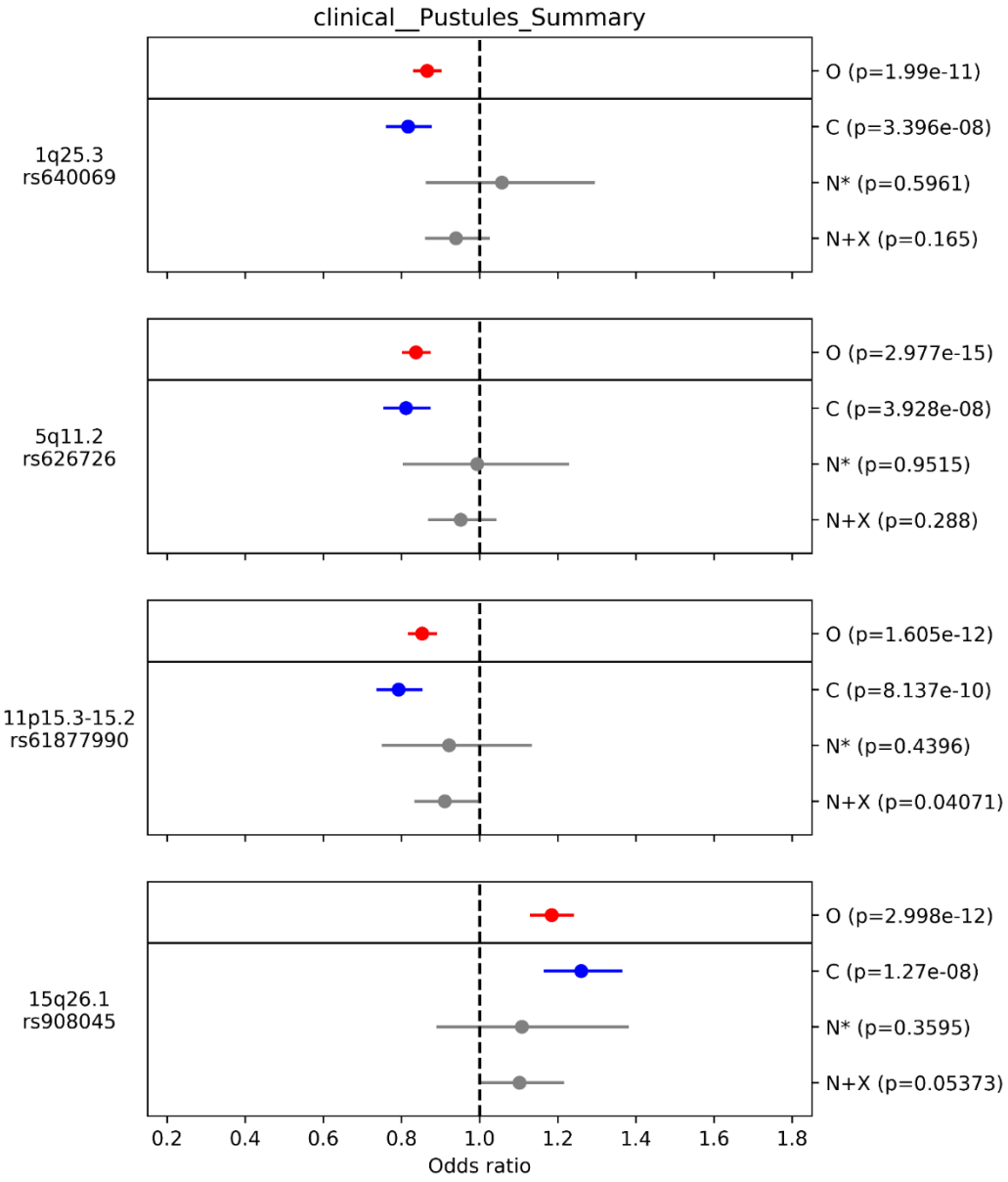


Figure 49: Loci where genome-wide significant levels of association were identified in individuals with pustule lesions. Odds ratios are shown for the SNP in both the original study meta-analysis (O; red) and each subphenotype analysis undertaken here (where genome-wide significant hits are shown in blue): C – using controls from original study; N – using "No" responses as controls; N+X – using non-"Yes" responses as controls. If there were insufficient numbers of individuals to conduct one arm of the meta-analyses, the summary statistics were reported for the single arm where numbers were sufficient (*).

Restricting the original meta-analysis to only individuals with cysts identified one of the previously discovered loci at genome-wide significance, but the odds ratio was not significantly than that of the original hit (Figure 50). Again, this signal was not replicated when comparing to individuals without cysts, whether stated explicity or including individuals with no response.

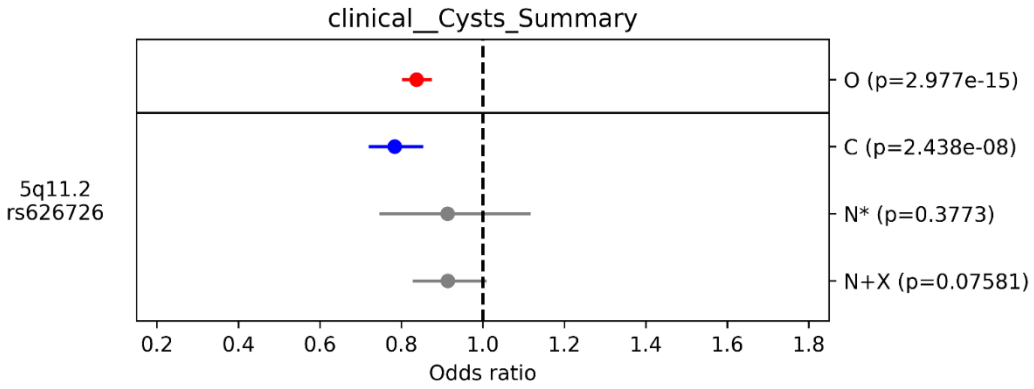


Figure 50: Locus where genome-wide significant levels of association were identified in individuals with cysts. Odds ratios are shown for the SNP in both the original study meta-analysis (O; red) and each subphenotype analysis undertaken here (where genome-wide significant hits are shown in blue): C – using controls from original study; N – using "No" responses as controls; N+X – using non-"Yes" responses as controls. If there were insufficient numbers of individuals to conduct one arm of the meta-analyses, the summary statistics were reported for the single arm where numbers were sufficient (*).

5.3.7 10q26.13 locus associated with cysts

Comparison of DS2 individuals with and without cysts identfied a genome-wide significant hit ($P = 1.34 \times 10^{-8}$) at a locus not previously described to be associated with acne (Figure 51). This locus was not signifiacntly associated with severe acne in the DS2 arm of the original meta-analysis in Petridis *et al*. ($P = 0.361$). When comparing individuals with cysts to unselected controls from the original analysis, the effect size is smaller than when comparing to acne individuals with confirmation of no cysts, though this falls short reaching genome-wide levels of statistical association ($P = 6.45 \times 10^{-3}$). The same is true when comparing to

individuals without cysts in which responses were interpreted as absence of cysts ($P = 1.60 \times 10^{-4}$).

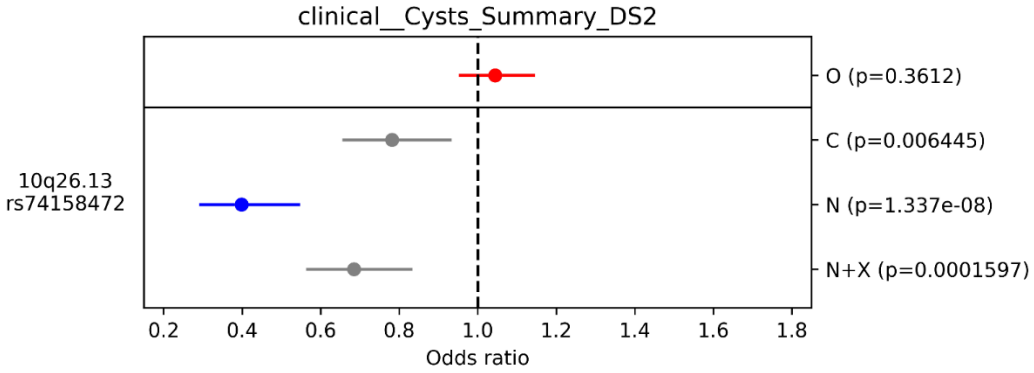


Figure 51: Novel locus where genome-wide significant levels of association were identified in individuals with cysts compared to individuals without cysts. Odds ratios are shown for the SNP in both the original study meta-analysis (O; red) and each subphenotype analysis undertaken here (where genome-wide significant hits are shown in blue): C – using controls from original study; N – using "No" responses as controls; N+X – using non-"Yes" responses as controls. Only the DS2 arm of the meta-analysis had sufficient individuals in the analysis that identified the genome-wide significant hit, so the summary statistics were reported for DS2 analyses only.

The protective allele is uncommon, with a frequency of 6.2% in 1142 individuals with cysts, 13.1% in 295 individuals where cysts were explicitly stated as absent, and 7.8% in the original healthy controls used for DS2. None of the genome-wide significant SNPs were directly genotyped but were imputed with an R^2 of at least 0.97. Association testing was only carried out on DS2 individuals because there were no individuals in DS1 for whom there was a "No" response for cysts. Further research is required to disentangle the mechanism through which variation at this locus may be involved in the cyst subphenotype. The only notable eQTLs associated with the lead SNP are in *ATE1* and *RP11-500G22.2* genes, differentially expressed in thyroid, lung, artery, nerve, adipose and transformed fibroblasts. The subphenotype-associated locus sits approximately 400 kb away from the *FGFR2* gene (Figure 52), the protein-product of which is a binding receptor for the *FGF2* ligand, which has been previously associated with acne through eQTL

colocalization. *FGF2* has established roles in wound healing and scarring (Petridis et al., 2018). Furthermore, rare mutations in *FGFR2* cause several bone dysplasia phenotypes.

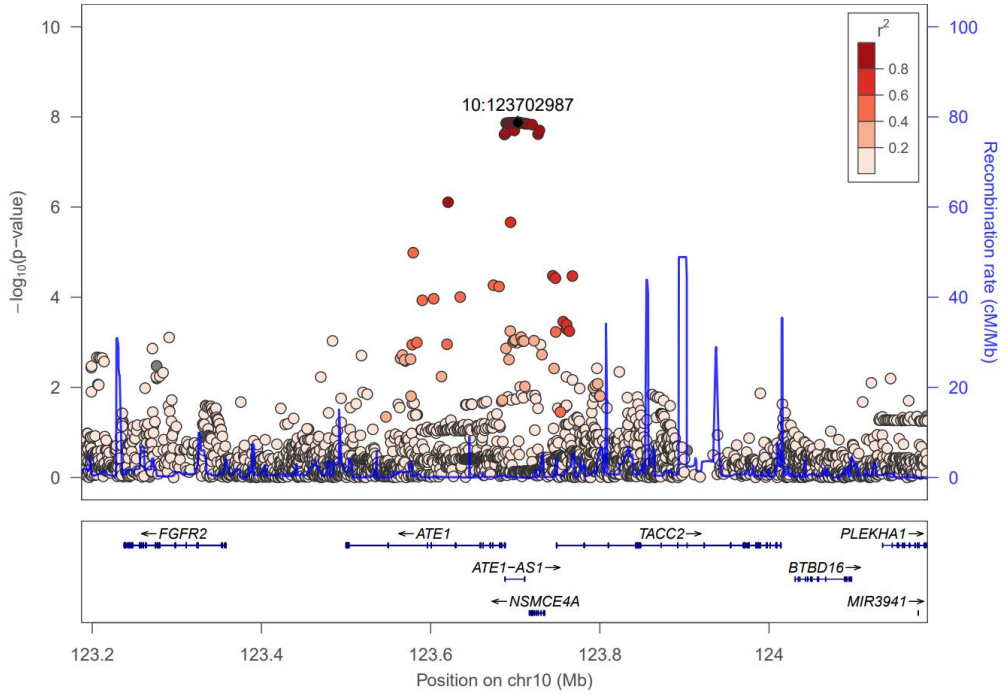


Figure 52: Association *P*-values of SNPs at the 10q26.13 locus for acne patients with cysts in DS2 (vs. DS2 individuals with acne and no cysts), indicating nearby genes and recombination rates of surrounding genomic locations.

5.4 Discussion

In this chapter phenotype data from acne patient questionnaires was used to identify suitable subphenotypes for genetic analysis. After filtering and encoding the data, applying dimensionality reduction and using density-based clustering, there were no robustly identified, interesting and sufficiently distinct subphenotypes with sufficient numbers for genetic analysis. Instead, subphenotypes were defined directly from question responses directly asked within the questionnaire, and those with sufficient numbers for genome-wide association analysis with a defined level of power were taken forward for analysis. Several loci previously associated with disease (Petridis et al., 2018) were associated with subphenotypes to genome-wide levels of significance, although there were no analyses of interest where subphenotypes were significantly more associated than the overall phenotype at any previous discovered loci. Analysis also identified one signal at a novel locus associated with cyst lesions, which warrants further investigation.

Collected data consisted of a CRF and a self-reported questionnaire for each patient. This contained a lot of extraneous information which required removal, such as basic patient/visit information, phenotypically irrelevant information and exclusion criteria. The data also contained potentially interesting question responses that existed in free-text, where computational encoding of the data for downstream analysis is not trivial. Furthermore, treatment information was also present, for which responder status would be of particular interest, but the section describing treatment information was generally very poorly filled (all questions had <2% response) and therefore was removed.

The questionnaires contained a large amount of missing information, and there was a large discrepancy between overall response rates between DS1 and DS2 due to

the presence of a subpopulation within DS2 with very highly filled questionnaires. This corresponded to negative responses in the scarring and lesion data. It is unclear how confidently the absence of a particular clinical trait can be imputed in a dataset of this nature, where most questions pertain to indication of presence of a particular clinical variant or lesion/scarring type. Here, blank responses in binary variables were assumed to be "No" responses for clustering and for one of the three arms of genetic analyses. However, they could also be interpreted as 'inconclusive', in which case a non-zero value could be argued to be more appropriate – the specific value may depend on the overall rarity of the trait within the dataset. Ultimately, the meaning of missing questionnaire data is dependent on the curation of the questionnaire, instructions given for filling them, and how these instructions are interpreted by individuals completing the questionnaire – therefore it can vary across different studies as well as within studies. It is important that such issues are considered when designing questionnaires and phenotype data models. Again, the same trade-off exists as with rare disease, where asking to fill deep phenotype information with maximal granularity will produce the most useful data for study but is more time-consuming and more likely to suffer from lack of compliance compared to more simple phenotype capture models.

The clustering methods employed a dimensionality reduction using t-SNE to calculate two-dimensional embedding representative of the distances between the high-dimensional data. The clustering didn't identify any sufficiently populous and distal phenotypic clusters which may have corresponded to an interesting subphenotype to study (e.g. the combination of a lesion type, scarring type and clinical variant, exclusively in males or females). Another potentially fruitful method of clustering includes combining questions (using either union or

intersection of "Yes" responses) based on disease knowledge (e.g. combination of nodulocystic acne and cyst lesion types into a single question) – the unsupervised dimensionality reduction approach was instead used it is potentially generalisable to other datasets without prior disease knowledge. One future direction of study is to analyse rare variants of individuals in this dataset, where identification of smaller phenotypic clusters would be more suitable.

Explicitly stated subphenotypes from questionnaire responses were instead used for genome-wide association analysis. Only binary variables were analysed, and spatial lesion and scarring data were not included because these were considered less clinically meaningful than the scarring/lesion types themselves. Potential quantitative traits such as disease severity scoring and age of onset were also not tested, which require further processing as none of them were normally distributed. Case-control analyses were conducted using different available genotyped cohorts as controls: (*i*) controls from the original studies (Navarini et al., 2014; Petridis et al., 2018) (*ii*) all individuals explicitly stated not to have the trait (*iii*) all individuals not stated to have the trait. Only analyses with 80% power to detect a common variant (50%) with a genotype relative risk of 2.0 at genome-wide levels of significance were conducted. Using the original control dataset as controls, 13 associations were identified at previously discovered loci in 6 subphenotypes, but none of the effect sizes were significantly larger than in the original GWAS meta-analysis apart from loci associated with cases with specified family history. No genome-wide associations were identified when defining controls as individuals within the acne cohort who weren't stated to have the subphenotype (i.e. considering blank data as "No" responses). However, when using explicit "Yes" and "No" responses as cases and controls, one locus without previous association

with acne was identified as associated with presence of cysts. There are eQTLs for nearby genes at this locus but neither the genes nor the expressed tissues are of particular phenotypic interest. The locus is near the *FGFR2* gene in which rare mutations cause several bone dysplasias, and a ligand of *FGFR2* (*FGF2*) has also been associated with acne through eQTL localisation (Petridis et al., 2018) – this warrants further investigation into the underlying mechanisms that drive association of this locus with only cysts but not the umbrella acne phenotype. The lack of other genome-wide associated signals may be due to a lack of statistical power in the analyses (thresholds were not stringent), or that the contribution of common genetic variation to the determination of acne subphenotypes is minimal. However, such methods may be applicable to other datasets where clinical information of a similar nature is recorded.

Chapter 6 – General Discussion

The first three results chapters in this thesis detail the development and application of methodologies for calculating phenotype similarity between a HPO term query and a reference set of diseases, including comparisons of performance in different clinical contexts. Computational identification of diseases similar to patient queries is important for the suggestion of differential clinical diagnoses, as well as for the identification of candidate genes in sequenced patients. Another important application of machine-readable phenotype similarity calculations is to identify subgroups of genotyped individuals with similar phenotypic presentations, something not tested in this thesis with rare disease, but with a dataset of acne patient questionnaires in the final results chapter. In this concluding chapter the main findings of the thesis are summarised and the key remaining challenges in these areas are discussed.

6.1 Benchmarking procedures

Benchmarking method performance began with simulations of queries against the OMIM reference phenotype set, where evaluation was based on the ranks of known similar phenotypes (Chapter 2). This was followed by querying HPO terms of patients from an external published dataset against the OMIM reference set, where evaluation was based on both the rank and score of the diagnostic gene (Chapter 3). This was repeated for diagnosed patients from a genetics clinic for whom whole exome sequencing data was available, evaluating based on rank and score of both diagnostic gene and variant (Chapter 4).

These benchmarking procedures compared several methodologies that form the composite process of estimating the phenotype similarity of a query to a reference

set. Firstly, the reference set of disease phenotypes requires characterisation with HPO terms, and here use of text mining was compared to manual curation – the utility of term quantification that can be derived from text mining was also tested against methods that use binary description. Secondly, machine-readable phenotypes must be algorithmically compared and their similarity quantified, and here vector-based cosine similarity was compared to $Resnik_{avg,max|sym}$, a similarity measure utilised within many published tools. Lastly, estimated probabilities conferred by use of phenotype similarity were compared to the implicit uniform probability distribution implied by the use of virtual gene panels.

6.1.1 OMIM phenotypic series

The first chapter made use of the OMIM phenotypic series (PS) as the group of known similar diseases and tested different combinations of phenotype annotation and similarity methods for their ability to group these diseases with each other rather than with other diseases within the OMIM catalogue. This enabled a large number and wide phenotypic range of diseases to be tested (which was a good representation of the full phenotypic spectrum of known human rare diseases; Figure 11, page 78) in method comparisons. This is in contrast to Phenomizer benchmarking where individual patients were simulated based on phenotypic features observed in 44 comprehensively documented dysmorphology syndromes (Köhler et al., 2009), where the quality of the benchmarking gold-standard was likely to be much higher, but tested on a much narrower phenotypic range of disease. The use of OMIM PS has further limitations, as evaluation is based on query similarity to other distinct disease entities within OMIM rather than the correct matching disease – although such disease entities have been judged to be clinically similar, this is only an approximation of the process of identifying

equivalent diseases that match the patient phenotypic constellation. However, these provided an estimation of the noise associated with patient queries, adding terms likely to occur in the relevant phenotypic constellations, which is usually simulated by adding randomly selected terms (Buske, Girdea, et al., 2015; Köhler et al., 2009). The use of the phenotypic series may also incur bias in the evaluation of both text-mined and curated reference sets: both the descriptive text and manually assigned HPO terms may be copied throughout diseases within the same OMIM phenotypic series. It is difficult to estimate the magnitude and effects of such bias. Due to these discussed drawbacks, OMIM phenotypic series benchmarking metrics were not sufficient alone to demonstrate superiority of any particular method of estimating phenotypic similarity. However, they demonstrated the potential of using text mining to characterise phenotypes and the use of the cosine rule to calculate similarity, particularly when HPO terms are quantified.

6.1.2 DDD

Benchmarking was then applied to a dataset of diagnosed patients from the DDD consortium study (The Deciphering Developmental Disorders Study, 2014; Wright et al., 2015). These represented a more homogenous group of patients (enriched for HPO terms within ‘Abnormality of the nervous system’, ‘Abnormality of head or neck’ and ‘Abnormality of the skeletal system’; Figure 17, page 100) due to it being a study of developmental disease, but the use of real patients emulates a genuine use case of the phenotype similarity methods. Similarity between patient queries and OMIM phenotypes was converted to gene similarity, and testing was based on both the rank and score of causative genes within the developmental disease virtual panel (DDG2P) that was used to diagnose patients. Testing the entire disease-associated genome would be more realistic in some respects, but only DDG2P

genes were tested here as diagnoses were made based on this gene panel (therefore it was not possible for other genes to feature). This benchmarking not only used on ranks of similar phenotypes (as in the OMIM PS benchmarking) for evaluation, but also considered scores of the phenotypic matches and how this influences the belief that certain genes were causative. Instead of using raw phenotype similarity scores, scores were rescaled with a logistic function, using similarity measured between phenotypes within the same OMIM PS – although this is a flawed solution (for reasons mentioned when discussing flaws in the OMIM PS benchmarking metrics), it was considered superior to using the arbitrary scales of different similarity scoring algorithms. These benchmarking metrics fall short of the ultimate use case of these similarity methods because they don't consider the sequence data of the phenotyped individuals – applying variant frequency and consequence filters to observed patient genetic variation dramatically reduces the number of potential variants (Figure 29, page 136), so it would be important to test how methods perform within this reduced search space.

6.1.3 Clinic diagnoses

Sequence data was available for diagnosed patients in the Guy's genetics clinic dataset, which enabled methods to be evaluated on the reduced search space derived by filtering observed genetic variation. However, due to the low numbers of patients with both high confidence diagnoses and assigned HPO terms there was low statistical power to compare methods. Only 31 of 65 patients with class 4 or 5 diagnostic variants were annotated with HPO terms. It is also not certain at which stage in the diagnostic process HPO terms were annotated, whereas the DDD stated that HPO term phenotyping was performed *before* diagnosis (Wright et al., 2015). Requesting HPO terms for the remaining diagnosed patients may be problematic

for benchmarking purposes because of the possibility that they would be biased by knowledge of the molecular diagnosis and the constellation of phenotypic features caused by the diagnostic gene.

6.1.4 Towards optimal benchmarking

It was important that benchmarking metrics moved from simulated queries to real patient queries and were ultimately used alongside patient sequence data. However, definitive conclusions could not be made due to a lack of sequenced individuals in the final comparison. It would be necessary to use another dataset – this would ideally involve requesting the DDD patient sequence data or using another equivalent dataset, although sufficiently large datasets with annotated HPO terms are difficult to curate. Benchmarking of variant prioritisation methods has commonly employed spiking disease-causing mutations from HGMD (Stenson et al., 2009) into exomes and simulating patient phenotypes (Buske, Girdea, et al., 2015; Javed et al., 2014; Peter N. Robinson et al., 2014; Smedley & Robinson, 2015; Zemojtel et al., 2014), whereas this thesis has focussed on using real-world data to inform gene prioritisation, as it contains the nuances of both the recorded phenotype information in clinics, as well as genotype data of rare disease patients.

A common theme in the benchmarking of phenotype similarity methods is the bias that may be incurred either as a result of the gold-standard dataset (i.e. OMIM curated and text-mined annotations affected by knowledge of the OMIM PS groupings), or during the formation of the gold-standard dataset (i.e. HPO annotations of patients influenced by knowledge of disorders on OMIM). Such effects are difficult to quantify and account for, so it is important when curating future benchmarking datasets to delineate where the HPO term annotation of patients sits within the diagnostic workflow.

It is also clear within datasets of diagnosed patients that the ease with which the diagnosis has been made is highly variable. Although methodologies developed here and elsewhere are designed to be capable of identifying straightforward diagnoses as well as offering candidates in hitherto unsolved cases, they are mostly required for the latter group, and it would be interesting to measure how different methods perform on such patient populations. It may be difficult to design such studies as it is not obvious how to measure the ease with which a diagnosis was made – potential proxies exist such as size of the prescribed virtual gene panel and time taken for diagnosis but these are not without clear limitations.

Though the range of disease phenotypes covered was reasonably homogeneous throughout benchmarking, it would be interesting to observe how methods perform on distinct disease areas, although large sample numbers for each area would be required for statistically powerful comparisons. Using the top HPO nodes below 'Phenotypic abnormality' is unlikely to be the optimal representation of phenotypic areas but it was the most feasible way to visualise disease phenotype categories across multiple datasets. It would also be of interest to measure how methods perform on syndromic phenotypes (where multiple organ systems are affected) compared to disease affecting a single organ system. The number of top HPO nodes below 'Phenotypic abnormality' represented within the annotated patient HPO terms would be a feasible measure of how "syndromic" a patient's disease phenotype is.

Considering all of the drawbacks mentioned, it is useful to imagine the "perfect" benchmarking evaluation dataset and whether it exists (or can exist). This would consist of sufficient numbers of real patients with high confidence genetic diagnoses (made using virtual gene panels) where HPO term assignment was made

prior to molecular diagnosis. It would also additionally be beneficial for these patients to be categorised into respective disease areas to observe differential method performance, and whether this is correspondent to diagnostic rates in the different disease areas. The 100,000 genomes rare disease programme is one such example of a large-scale sequencing project where patients have been recruited in broad categories (GeCIP domains). Roughly 17,000 individuals with rare diseases will be sequenced (Genomics England, n.d.), and with an initial diagnostic rate of 22% (Turnbull et al., 2018), an estimated 3,740 rare disease patients in total would be available for benchmarking. However, this again may contain some bias due to the increasing number of phenotype-driven tools that are used to aid diagnosis in such projects.

6.2 Methodology comparisons

6.2.1 Curation vs. text mining

One methodology comparison undertaken in this thesis is the utility of text mining for phenotypes compared to manual curation. It is important to gauge the efficacy of text mining for phenotypes because machine-readable phenotype annotation is still not yet universally adopted. Even in instances where HPO annotations have been made, the diligence with which it is performed (and therefore the quality and depth of phenotyping) can be highly variable, so the use of text mining free-text descriptions can add value (especially as natural language processing methods become increasingly sophisticated). Without quantification of terms, text-mining of phenotypes was slightly inferior to manual curation in the OMIM PS benchmarking (Figure 12, page 81), and there was no significant difference in ranking causative genes in the DDD and clinic benchmarking (Figure 18, page 103; Figure 25, page 130; Figure 30, page 139).

Text mining also identified HPO terms in a subset of 14 individuals from the Guy's genetics clinic, two of which had likely diagnostic variants. Both diagnostic variants were identified in first place when ranking genes by gene phenotype similarity to the text-mined HPO terms, and additionally, an interesting group of potential diagnostic variants (also ranked in first place) were identified in another individual, which may not have otherwise been identified.

The OMIM reference phenotype annotation set used here was text-mined in February 2016, which was compared to the curated reference set downloaded at the same time to enable fair comparison between methods. These same annotation sets were compared throughout the benchmarking procedures, and the quantified text-mined reference set was used for HPO queries of the undiagnosed clinic patients – however, suggestion of diagnostic variants would be most effective if the most up-to-date version of the reference set could be used.

6.2.2 Quantification vs. no quantification

Testing also compared HPO term quantification using text-mining, which should allow more relevant or important terms to be up-weighted in subsequent analysis, against methods that don't incorporate quantification. Quantified text-mined phenotypes were tested against manually curated phenotypes (where quantification information wasn't used), as well as an unquantified version of the same text-mined set. In the OMIM PS benchmarking, the use of quantification of reference phenotypes greatly increased the ability to group relevant phenotypes together, though this was less clear when the query terms themselves were not quantified (Figure 12, page 81). In the DDD benchmarking, quantification of the phenotype reference set was significantly better at ranking patient causative genes (Figure 18, page 103), though there was no evidence of this in the clinic dataset (Figure 25,

page 130; Figure 30, page 139). Text-mining was limited to only the OMIM free-text description, but could be expanded, for example to source publications cited in OMIM. Further text-mining would refine term frequencies based on relevance, though it would also introduce much more noise. Future work should include testing between text-mined and curated quantification, though this will require encoding the many missing quantification values in the curated dataset.

6.2.3 Cosine vs. Resnik

For calculation of similarity between a query and reference set of disease phenotypes, vector-based cosine similarity was compared to $Resnik_{avg,max|sym}$, which was selected for comparison with vector-based cosine similarity due to its implementation in several tools (Javed et al., 2014; Köhler et al., 2009; Zemojtel et al., 2014). Vector-based similarity showed improvements ahead of the node-based Resnik, particularly when the reference set was quantified (Figure 12, page 81; Figure 18, page 103). *P*-value calculation was not incorporated in comparisons, which has been shown to be beneficial when imprecision is simulated (when imprecision and noise have been simulated separately) (Köhler et al., 2009). Incorporation of *p*-value estimation is likely to increase performance of any similarity measure and theoretically could be applied to any similarity method (including cosine similarity). *P*-value estimation considers the local score distribution of each reference phenotype, which is more specific than the logistic function developed which only considers the global reference set score distribution.

6.2.4 BOQA and PhenIX

Other published tools for suggestion of clinical diagnostics and variant prioritisation were compared to the methods investigated throughout this thesis. BOQA (Bauer et al., 2012) was compared within the DDD benchmarking, where

it was inferior in ranking disease genes (Figure 18, page 103), though it was able to identify a subset of the diagnoses with high confidence (Figure 19, page 106). When BOQA incorporated the reference disease phenotype frequency information (from both curated and text mining annotation), it did not produce significantly improved results. PhenIX (Zemojtel et al., 2014) was used for comparison in the variant prioritisation benchmarking of clinic patients (Figure 30, page 139), although this comparison did not have sufficient power to draw any firm conclusion.

6.2.5 Comparisons omitted

There are a number of phenotype similarity methods that weren't compared to those developed in this thesis, such as PhenoDigm (Smedley et al., 2013) and simGIC (Pesquita et al., 2007) which have performed notably well in other comparisons (Buske, Girdea, et al., 2015). $Resnik_{avg,max|sym}$ was selected for comparison as it underlies PhenIX, which has been demonstrated in multiple reviews to be superior at phenotype-based prioritisation of variants in known human disease genes (Pengelly et al., 2017; Smedley & Robinson, 2015). Ideally the maximal number of published methods would be compared in these benchmarking tests, though here the strengths and weaknesses of the various steps involved in calculating phenotype similarity were assessed, a non-trivial comparison to which the addition of further published tools would add limited further insight.

6.3 Considerations for all machine-readable phenotype similarity methodology

6.3.1 Adding value to approaches that don't utilise machine-readable phenotypes

The utility of this methodology can firstly be demonstrated in the comparison to the use of the DDG2P panel in the DDD patient testing, where patient diagnostic genes are almost always ranked in the top 500 of the 3,303 gene panel. The diagnostic gene was prioritised within the top 100 in 56% of cases and the top 10 in 23% of cases by the top-performing method (quantified text mining, cosine similarity; Figure 18, page 103). Furthermore, for the majority of methods the probability-rescaled scores of diagnostic genes were almost always superior to randomly selecting the gene from the panel. In the clinic patients there were superior statistics for the causative gene appearing in the top 10 and 100 disease-associated genes were observed (34% and 69% respectively; Figure 25, page 130), but it remains to be seen how this benefit replicates upon observing patient genetic variation. The prioritised disease-associated genome was not compared to the use of gene panels, due to the highly variable size of gene panels (and resultant numbers of variants post-filtering), which was further complicated because the diagnostic gene wasn't contained within some of the prescribed virtual gene panels.

Further utility of this methodology is demonstrated with the suggestion of interesting variants in cases for which no answer has been found (though even if such suggested variants are confirmed, it is difficult to quantify this benefit in comparison to the standard operating procedure in diagnostic labs which can be variable between both labs and individuals within labs performing variant analysis). However, it is important to note that the incorporation of phenotype similarity metrics to rank/filter/suggest variants requires minimal effort other than

the annotation of patients with HPO terms (which is becoming increasingly adopted). It is also uncertain whether the suggested variants listed here have been highlighted or scrutinised previously, and whether evidence from phenotype similarity metrics would lend weight to assessment of pathogenicity, which is something not addressed in ACMG guidelines (which does consider other scoring such as tolerance/deleteriousness predictions). The use of text mining clinic letters also enabled the identification of possible compound heterozygous variants which are seemingly deleterious and within a gene highly relevant to the patient phenotype. HPO terms were not supplied for the patient and the gene was not in the prescribed gene panel so without these methods this variant may not have otherwise been identified. Text mining clinic letters for phenotypic information theoretically requires little additional clinical effort and can detect otherwise uncaptured HPO terms for the suggestion of potential diagnostic variants. However, in practice here it required manual anonymisation of clinic letters before transfer into a research environment (hence the low numbers of individuals for which this was performed), but this could be circumvented by including HPO text mining tools within installed software approved for use in diagnostic labs.

6.3.2 Missing phenotype data

In the limited sample of patients from the genetics clinic it was interesting to observe that only half of patients were annotated with HPO terms, and that the proportion was similar amongst diagnosed and undiagnosed individuals. Although machine-readable phenotypes are expected to be increasingly recorded through data models instituted by large rare disease sequencing projects (Caulfield et al., 2015; The Deciphering Developmental Disorders Study, 2014), there is a need for methods that can identify phenotype information where it is not provided. The

potential of text mining health records has been demonstrated here, though it still requires rigorous testing against high-standard manual curation to properly assess the sensitivity and specificity with which it identifies HPO terms. HPO terms could also be identified using the latent phenotypic information contained in prescribed virtual gene panels by working backward from gene to OMIM disease, and then extracting HPO terms (potentially scoring them based on the proportion of gene-phenotype pairs from the gene list that they feature in). Another interesting question is whether HPO annotations can be assessed for informative-ness among individuals to prioritise patients for sequencing (in highly informative cases) or further annotation (in non-informative cases). Such measures of informative-ness are available in PhenoTips data entry (via the Monarch phenotype specificity meter), though this only evaluates queries on the information content of their terms (Girdea et al., 2013). The quality of HPO content could be assessed based on the distribution of similarity scores between an individual's annotations and the reference set, or the number of genes with a significant phenotypic match.

The use of questionnaires can encourage more complete curation of clinical information by prompting the user to input data about specific phenotypic characteristics, though the application of such data models can only be instituted within a specific disease area (otherwise questionnaire options would be almost limitless). This method of phenotypic data collection was performed for patients with severe acne in cohorts from two genetic studies (Navarini et al., 2014; Petridis et al., 2018), with records available for 98% of individuals contained in the final analysis of these studies. These records contained large amounts of missing data, and it is unclear whether missing data should be interpreted as absent or inconclusive, and how such inferences should be numerically encoded. When

applying clustering methods missing data for binary questions was interpreted as absence of a particular clinical feature or clinical variant of acne and missing quantitative measures were replaced with median values. Additionally, one arm of the genetic association testing described in chapter 5 used individuals within the cohort with missing data for a particular question as controls, for which there is much less confidence in true control status than individuals where absence is explicitly stated. These solutions are not ideal, and could potentially be improved for specific questions: models could be trained to predict quantitative values (e.g. acne Leeds score) based on the type(s) and location(s) of lesions and scarring, and vice-versa. However, considering the overall response rate of the questionnaires, sample sizes would presumably be prohibitively low to train robust predictors.

Future work could consider extensions to the standard HPO term annotation of a phenotype. Negative rare disease annotations were not considered by methods tested in this thesis (though negative annotations were made in many clinic patients due to functionality of the PhenoTips interface) – here all HPO terms not present were assumed to be absent. False negative probabilities can be estimated for HPO terms not contained in the query (which haven't been negatively annotated), which is attempted in BOQA using uniform global levels of false negative probability for query terms (Bauer et al., 2012). Phenotypes could be "imputed" more specifically using observed data documenting co-occurring phenotypic features, but this would require a large and richly phenotyped dataset to make such predictions. This is further complicated as syndromic phenotypes in particular are not 'static' and different phenotypic features manifest at different timepoints in development. HPO clinical modifier terms exist for age of onset, pace of progression, phenotypic variability and positions among others, but these are not utilised by methods tested

in this thesis. Theoretically clinical modifier terms can exist as subnodes for every phenotypic term within the ontology, where maximal similarity is conferred in instances where both the term and its clinical modifier match (though such a model may experience difficulties with implicit annotation of ancestor terms). However, such data is not frequently annotated to the reference set (611 of 152,014 annotations in the most recent curated HPO annotation set contain onset modifiers), and so would require a tremendous curation effort or sophisticated text mining and natural language processing techniques to associate such modifiers.

6.3.3 Phenotypic clusters

Based on severe acne patient questionnaire data, dimensionality reduction and density-based clustering methods were used to identify phenotypic clusters corresponding to putative acne subphenotypes. This was unsuccessful, with the only large and distinct clusters corresponding to individuals with high amounts of missing data, which separated based on the remaining data fields into sex and indication of the most common clinical variant of acne contained in the study, nodulocystic acne. It is possible that a genetically determined subtype of acne lay within smaller phenotypic clusters, but statistical power to identify genome-wide-associated common SNPs is limited in such cases. Identification of phenotypic clusters was not attempted in the clinic patients – there were only 55 undiagnosed patients with HPO terms available (69 including those with VUS; Table 10, page 119), which existed on a wide phenotype spectrum, causing the data to be sparse. It may be possible to identify small phenotypic clusters for rare disease using visualisation methods described above but these would likely correspond to the sharing of non-specific terms. Such methods may be more likely to work in larger

datasets, such as the undiagnosed DDD patients, in which the range of disease phenotypes would also be narrower than a genetics clinic.

6.3.4 Estimating statistical significance of measures of phenotype similarity

Scoring similarity between individual's phenotype and a reference set, or other individuals' phenotypes benefits greatly from consideration of the statistical significance of certain levels of similarity score. Methods developed in this thesis employ global empirical similarity score distributions in similar and non-similar phenotypes based on phenotypic series data to estimate the probability associated with observing a given level of similarity score. Phenomizer (and Phenomizer-incorporating) methods utilise an alternative approach of assessing local similarity scores for each of the reference phenotypes by generating random searches for differing numbers of query terms (Köhler et al., 2009), and similar approaches may be suitable for assessing the complete probability distribution of patient HPO term similarity through random searches sampled on the relative frequencies of each phenotypic term within large rare disease patient datasets.

6.3.5 Coding and non-coding variation

The methodologies tested and applied here only consider variation within the coding regions of the genome. It is relatively straightforward to establish causality of a novel coding variant within a known disease gene through functional consequence and prediction data, compared to establishing the causality of the non-coding variant. However, the frequency with which non-coding variants cause disease is unknown because disease mechanisms are often less clear and have been the subject of much fewer studies. The low diagnostic rate of rare disorders in general makes non-coding variation a potentially fruitful area of study. CADD scores are able to evaluate non-coding SNVs (Kircher et al., 2014), and variant

analysis frameworks (including additional scoring of non-coding positions) are available for the inclusion of non-coding variants in phenotype-aware variant prioritisers that combine phenotype and variant score (Smedley et al., 2016). However, in general, much progress will be need in the understanding of non-coding causes of rare disease for the methods developed in this thesis to become directly applicable.

6.3.6 Combining variant scoring with phenotype-based prioritisation

In this thesis patient exome variants were filtered prior to phenotype-based prioritisation based on functional consequence (removal of synonymous variation) and estimates of allele frequency, and such information was not additionally used to prioritise variants (i.e. preferentially prioritising nonsense and frameshift mutations over missense). Pathogenicity prediction scores were also not used to filter variants or influence the prioritisation of post-filtering variants. Several published methods combine variant-level scoring (using combinations of variant frequency, consequence and pathogenicity prediction tools) with gene-level scoring (based on phenotypic relevance) (Bone et al., 2015; Javed et al., 2014; Sifrim et al., 2013; Singleton et al., 2014; Zemojtel et al., 2014). Addition of variant-level information is expected to increase performance of phenotype-based methods, as shown in benchmarking of these tools. Furthermore, mode of inheritance information can also be incorporated into variant filtering – firstly, genes can be removed where the patient genotype is not consistent with the known phenotype-gene relationships, and secondly, where there is a suspected mode of inheritance in the patient, this can be used to filter/prioritise variants based on whether the genotypes are consistent. Such additions also have potential to improve prioritisation performance, and would be an interesting direction of future study.

6.4 Conclusion

This thesis has demonstrated the utility of text mining in annotating rare disease phenotypes for both reference diseases and patient phenotypic descriptions. Novel benchmarking strategies have been employed using similar phenotypes defined within the reference set, as well as a group of diagnosed individuals in a publicly available dataset. Methods were also tested in the clinic, although the limited number of patients lowered the power of statistical comparison of methods. They were also applied to undiagnosed patients in the clinic, suggesting interesting candidate variants which warrant additional investigation. Lastly, questionnaire data for individuals with a common complex disease (acne) was used to identify subphenotypes, and despite the dataset containing large amounts of missing data, subsequent genetic analysis identified a genome-wide significant hit in a subphenotype which warrants further investigation. Taken together, this work has therefore demonstrated that careful utilisation of phenotype data has great potential to aid genetic diagnosis and discovery, and that further research in this area should prove both interesting and informative.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*. https://doi.org/10.1038/nature15393

Adegbola, A., Musante, L., Callewaert, B., Maciel, P., Hu, H., Isidor, B., … Kalscheuer, V. M. (2015). Redefining the MED13L syndrome. *European Journal of Human Genetics*, (January), 1–10. https://doi.org/10.1038/ejhg.2015.26

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, (SUPPL.76). https://doi.org/10.1002/0471142905.hg0720s76

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., … Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*. https://doi.org/10.1038/nbt1203

Al Olama, A. A., Kote-Jarai, Z., Berndt, S. I., Conti, D. V., Schumacher, F., Han, Y., … Haiman, C. A. (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nature Genetics*. https://doi.org/10.1038/ng.3094

Alster, T. S., & West, T. B. (1997). Treatment of scars: A review. *Annals of Plastic Surgery*. https://doi.org/10.1097/00000637-199710000-00014

Amberger, J. S., Bocchini, C. a, Schiettecatte, F., Scott, A. F., & Hamosh, A. (2014). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*(November 2014), 789–798. https://doi.org/10.1093/nar/gku1205

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*. https://doi.org/10.1038/75556

Australian Digital Health Agency. (n.d.). Clinical Terminology. Retrieved June 14, 2018, from https://www.digitalhealth.gov.au/get-started-with-digital-health/what-is-digital-health/clinical-terminology

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3031

Bard, J., Rhee, S. Y., & Ashburner, M. (2005). An ontology for cell types. *Genome Biology*. https://doi.org/10.1186/gb-2005-6-2-r21

Bataille, V., Snieder, H., MacGregor, A. J., Sasieni, P., & Spector, T. D. (2002). The influence of genetics and environmental factors in the pathogenesis of acne: A twin study of acne in women. *Journal of Investigative Dermatology*. https://doi.org/10.1046/j.1523-1747.2002.19621.x

Bauer, S., Köhler, S., Schulz, M. H., & Robinson, P. N. (2012). Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics, 28*(19), 2502–2508. https://doi.org/10.1093/bioinformatics/bts471

Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., … Boycott, K. M. (2014). FORGE Canada consortium: Outcomes of a 2-year national rare-disease gene-discovery project. *American Journal of Human Genetics*, *94*(6), 809–817. https://doi.org/10.1016/j.ajhg.2014.05.003

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Blekhman, R., Man, O., Herrmann, L., Boyko, A. R., Indap, A., Kosiol, C., … Przeworski, M. (2008). Natural Selection on Genes that Underlie Human Disease Susceptibility. *Current Biology*. https://doi.org/10.1016/j.cub.2008.04.074

Bone, W. P., Washington, N. L., Buske, O. J., Adams, D. R., Davis, J., Draper, D., … Smedley, D. (2015). Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine*. https://doi.org/10.1038/gim.2015.137

Boycott, K. M., Dyment, D. A., Sawyer, S. L., Vanstone, M. R., & Beaulieu, C. L. (2014). Identification of Genes for Childhood Heritable Diseases. *Annual Review of Medicine*. https://doi.org/10.1146/annurev-med-101712-122108

Burke, B. M., & Cunliffe, W. J. (1984). The assessment of acne vulgaris—the Leeds technique. *British Journal of Dermatology*. https://doi.org/10.1111/j.1365-2133.1984.tb04020.x

Buske, O. J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., … Brudno, M. (2015). PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. *Human Mutation*. https://doi.org/10.1002/humu.22851

Buske, O. J., Schiettecatte, F., Hutton, B., Dumitriu, S., Misyura, A., Huang, L., … Brudno, M. (2015). The Matchmaker Exchange API: Automating Patient Matching Through the Exchange of Structured Phenotypic and Genotypic Profiles. *Human Mutation*. https://doi.org/10.1002/humu.22850

Canada Health Infoway. (n.d.). SNOMED CT. Retrieved June 14, 2018, from https://infocentral.infoway-inforoute.ca/en/standards/international/snomed-ct

Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., … Woods, K. (2015). The 100,000 Genomes Project Protocol. *Genomics England*. https://doi.org/10.6084/M9.FIGSHARE.4530893.V2

Chatzimichali, E. A., Brent, S., Hutton, B., Perrett, D., Wright, C. F., Bevan, A. P., … Swaminathan, G. J. (2015). Facilitating Collaboration in Rare Genetic Disorders Through Effective Matchmaking in DECIPHER. *Human Mutation*. https://doi.org/10.1002/humu.22842

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., … Lifton, R.

P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(45), 19096–19101. https://doi.org/10.1073/pnas.0910672106

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*. https://doi.org/10.1038/ng.3656

de Ligt, J., Willemsen, M. H., van Bon, B. W. M., Kleefstra, T., Yntema, H. G., Kroes, T., … Vissers, L. E. L. M. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *The New England Journal of Medicine*, *367*(20), 1921–1929. https://doi.org/10.1056/NEJMoa1206524

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., … Roden, D. M. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*. https://doi.org/10.1038/nbt.2749

Denny, J. C., Crawford, D. C., Ritchie, M. D., Bielinski, S. J., Basford, M. A., Bradford, Y., … De Andrade, M. (2011). Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: Using electronic medical records for genome- and phenome-wide studies. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2011.09.008

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., … Crawford, D. C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq126

Department of Health, & UK Government. (2013). *The UK Strategy for Rare Diseases*. *Department of Health and Social Care Publication*.

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. https://doi.org/10.1038/ng.806

Dewey, F. E., Murray, M. F., Overton, J. D., Habegger, L., Leader, J. B., Fetterolf, S. N., … Carey, D. J. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*. https://doi.org/10.1126/science.aaf6814

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2809

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*. https://doi.org/10.1186/gb-2005-6-5-r44

Ellard, S., Baple, E. L., Owens, M., Eccles, D. M., Abbs, S., & Zandra, C. (2017). ACGS Best Practice Guidelines for Variant Classification 2017. *Association*

for Clinical Genetic Science, 1–12.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2005.10.010

Finlay, A. Y., & Khan, G. K. (1994). Dermatology Life Quality Index (DLQI) - A simple practical measure for routine clinical use. *Clinical and Experimental Dermatology*. https://doi.org/10.1111/j.1365-2230.1994.tb01167.x

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., … Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, *84*(4), 524–533. https://doi.org/10.1016/j.ajhg.2009.03.010

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., … Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. https://doi.org/10.1038/nature11690

Gahl, W. A., Mulvihill, J. J., Toro, C., Markello, T. C., Wise, A. L., Ramoni, R. B., … Tifft, C. J. (2016). The NIH Undiagnosed Diseases Program and Network: Applications to modern medicine. *Molecular Genetics and Metabolism*. https://doi.org/10.1016/j.ymgme.2016.01.007

Genomics England. (n.d.). The 100,000 genomes project.

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3118

Gilissen, C., Arts, H. H., Hoischen, A., Spruijt, L., Mans, D. A., Arts, P., … Brunner, H. G. (2010). Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2010.08.004

Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K. M., Chénier, S., … Brudno, M. (2013). PhenoTips: Patient phenotyping software for clinical and research use. *Human Mutation*, *34*(8), 1057–1065. https://doi.org/10.1002/humu.22347

Goldstein, D. B., Allen, A., Keebler, J., Margulies, E. H., Petrou, S., Petrovski, S., & Sunyaev, S. (2013). Sequencing studies in human genetics: Design and interpretation. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3455

Guarino, N. (1998). Formal Ontology and Information Systems. In *FOIS'98* (Vol. 46, pp. 3–15). https://doi.org/10.1.1.29.1776

Hahm, B. J., Min, S. U., Yoon, M. Y., Shin, Y. W., Kim, J. S., Jung, J. Y., & Suh, D. H. (2009). Changes of psychiatric parameters and their relationships by oral isotretinoin in acne patients. *Journal of Dermatology*. https://doi.org/10.1111/j.1346-8138.2009.00635.x

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., … Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1031

He, L., Wu, W. J., Yang, J. K., Cheng, H., Zuo, X. B., Lai, W., … Zhang, Y. P.

(2014). Two new susceptibility loci 1q24.2 and 11p11.2 confer risk to severe acne. *Nature Communications*. https://doi.org/10.1038/ncomms3870

Hoischen, A., van Bon, B. W. M., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., … Veltman, J. A. (2010). De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature Genetics*, *42*(6), 483–485. https://doi.org/10.1038/ng.581

Howard, D. M., Adams, M. J., Shirali, M., Clarke, T. K., Marioni, R. E., Davies, G., … McIntosh, A. M. (2018). Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nature Communications*. https://doi.org/10.1038/s41467-018-03819-3

IHTSDO. (2014). SNOMED CT Starter Guide. *SNOMED*, 1–56.

Institute of Medicine (US) Committee on Accelerating Rare Diseases Research and Orphan Product Development. (2010). *Rare Diseases and Orphan Products: Accelerating Research and Development*. (M. Field & T. Boat, Eds.). Washington (DC): National Academies Press (US).

Jacob, C. I., Dover, J. S., & Kaminer, M. S. (2001). Acne scarring: A classification system and review of treatment options. *Journal of the American Academy of Dermatology*. https://doi.org/10.1067/mjd.2001.113451

Javed, A., Agrawal, S., & Ng, P. C. (2014). Phen-gen: Combining phenotype and genotype to analyze rare disorders. *Nature Methods*. https://doi.org/10.1038/nmeth.3046

Jiang, J., & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the 10th International Conference on Research on Computational Linguistics, Taiwan*.

Kaiser, J. (2010). Affordable "exomes" fill gaps in a catalog of rare diseases. *Science*. https://doi.org/10.1126/science.330.6006.903

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., … Schriml, L. M. (2015). Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku1011

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., Shendure, J., … Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., … Robinson, P. N. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, *42*(Database issue), D966-74. https://doi.org/10.1093/nar/gkt1026

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., … Robinson, P. N. (2009). Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *American Journal of Human Genetics*, *85*, 457–464. https://doi.org/10.1016/j.ajhg.2009.09.003

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., … Holmen, J. (2013). Cohort profile: The HUNT study, Norway. *International Journal of Epidemiology*. https://doi.org/10.1093/ije/dys095

Kubota, Y., Shirahige, Y., Nakai, K., Katsuura, J., Moriue, T., & Yoneda, K. (2010). Community-based epidemiological study of psychosocial effects of acne in Japanese adolescents. *Journal of Dermatology*. https://doi.org/10.1111/j.1346-8138.2010.00855.x

Kuhlenbäumer, G., Hullmann, J., & Appenzeller, S. (2011). Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Human Mutation*. https://doi.org/10.1002/humu.21400

Kulminski, A. M., Loika, Y., Culminskaya, I., Arbeev, K. G., Ukraintseva, S. V., Stallard, E., & Yashin, A. I. (2016). Explicating heterogeneity of complex traits has strong potential for improving GWAS efficiency. *Scientific Reports*. https://doi.org/10.1038/srep35390

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1081. https://doi.org/10.1038/nprot.2009.86

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., … Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, *25*(3), 309–316. https://doi.org/10.1038/nbt1295

Lalonde, E., Albrecht, S., Ha, K. C. H., Jacob, K., Bolduc, N., Polychronakos, C., … Jabado, N. (2010). Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Human Mutation*. https://doi.org/10.1002/humu.21293

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx1153

Lee, D., Cornet, R., Lau, F., & de Keizer, N. (2013). A survey of SNOMED CT implementations. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2012.09.006

Lee, H., Deignan, J. L., Dorrani, N., Strom, S. P., Kantarci, S., Quintero-Rivera, F., … Nelson, S. F. (2014). Clinical exome sequencing for genetic identification of rare mendelian disorders. *JAMA - Journal of the American Medical Association*. https://doi.org/10.1001/jama.2014.14604

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2014.06.009

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. https://doi.org/10.1038/nature19057

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin,

R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp352

Li, M. X., Yeung, J. M. Y., Cherny, S. S., & Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics*. https://doi.org/10.1007/s00439-011-1118-2

Lin, D. (1998). An Information-Theoretic De nition of Similarity. In *Proceedings of the 15th International Conference on Machine Learning*.

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., … Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. https://doi.org/10.1126/science.1215040

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., … Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*. https://doi.org/10.1038/nature13127

MacRae, C. A., & Vasan, R. S. (2011). Next-Generation Genome-Wide Association Studies: Time to Focus on Phenotype? *Circ Cardiovasc Genet.*, *4*(4), 334–336.

Maiella, S., Rath, A., Angin, C., Mousson, F., & Kremp, O. (2013). Orphanet and its consortium: Where to find expert-validated information on rare diseases. *Revue Neurologique*. https://doi.org/10.1016/S0035-3787(13)70052-3

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., & Jabado, N. (2011). What can exome sequencing do for you? *Journal of Medical Genetics*. https://doi.org/10.1136/jmedgenet-2011-100223

Manning, C. D., & Raghavan, P. (2009). An Introduction to Information Retrieval. In *Online*. https://doi.org/10.1109/LPT.2009.2020494

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2796

McKusick, V. A. (1966). *Mendelian Inheritance in Man*. *Mendelian Inheritance in Man*. https://doi.org/10.1016/B978-1-4831-6679-7.50008-4

McKusick, V. A. (2007). Mendelian Inheritance in Man and Its Online Version, OMIM. *The American Journal of Human Genetics*. https://doi.org/10.1086/514346

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2626

Ministry of Health - New Zealand Government. (2017). SNOMED CT. Retrieved June 14, 2018, from https://www.health.govt.nz/nz-health-statistics/classification-and-terminology/new-zealand-snomed-ct-national-release-centre/snomed-ct

Mitchell, J. S., Li, N., Weinhold, N., Försti, A., Ali, M., Van Duin, M., … Houlston,

R. S. (2016). Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. *Nature Communications*. https://doi.org/10.1038/ncomms12050

Muncke, N., Jung, C., Rüdiger, H., Ulmer, H., Roeth, R., Hubert, A., … Rappold, G. (2003). Missense Mutations and Gene Interruption in PROSIT240, a Novel TRAP240-Like Gene, in Patients with Congenital Heart Defect (Transposition of the Great Arteries). *Circulation*, *108*(23), 2843–2850. https://doi.org/10.1161/01.CIR.0000103684.77636.CD

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M.-A., & Smith, B. (2012). The National Center for Biomedical Ontology. *Journal of the American Medical Informatics Association*, *19*(2), 190–195. https://doi.org/10.1136/amiajnl-2011-000523

Musunuru, K., Pirruccello, J. P., Do, R., Peloso, G. M., Guiducci, C., Sougnez, C., … Kathiresan, S. (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *The New England Journal of Medicine*, *363*(23), 2220–2227. https://doi.org/10.1056/NEJMoa1002926

Navarini, A. A., Simpson, M. A., Weale, M., Knight, J., Carlavan, I., Reiniche, P., … Barker, J. N. (2014). Genome-wide association study identifies three novel susceptibility loci for severe Acne vulgaris. *Nature Communications*. https://doi.org/10.1038/ncomms5020

Neveling, K., Feenstra, I., Gilissen, C., Hoefsloot, L. H., Kamsteeg, E. J., Mensenkamp, A. R., … Nelen, M. R. (2013). A Post-Hoc Comparison of the Utility of Sanger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases. *Human Mutation*. https://doi.org/10.1002/humu.22450

Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., … Shendure, J. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics*. https://doi.org/10.1038/ng.646

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., … Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, *42*(1), 30–35. https://doi.org/10.1038/ng.499

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., … Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, *461*(7261), 272–276. https://doi.org/10.1038/nature08250

NHS. (n.d.). SNOMED CT. Retrieved June 14, 2018, from https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct

Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., … Farrall, M. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*. https://doi.org/10.1038/ng.3396

Novocraft. (2014). Novocraft. Retrieved August 29, 2018, from

http://www.novocraft.com/.

Pengelly, R. J., Alom, T., Zhang, Z., Hunt, D., Ennis, S., & Collins, A. (2017). Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. *Scientific Reports*. https://doi.org/10.1038/s41598-017-13841-y

Pesquita, C., Faria, D., Bastos, H., Falcão, A. O., & Couto, F. M. (2007). Evaluating GO-based semantic similarity measures. *InProceedings of 10thAnnualBio-OntologiesMeeting*.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. *PLoS Comput Biol*, *5*, e1000443. https://doi.org/10.1371/journal.pcbi.1000443

Petridis, C., Navarini, A. A., Dand, N., Saklatvala, J., Baudry, D., Duckworth, M., … Simpson, M. A. (2018). Genome-wide association study and meta-analysis implicates mediators of hair follicle development and morphogenesis as risk factors for severe acne. *In Press*.

Purdy, S., & de Berker, D. (2011). Acne vulgaris. *BMJ Clinical Evidence*. https://doi.org/1714 [pii]

Quang, D., Chen, Y., & Xie, X. (2015). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu703

Rath, A., Olry, A., Dhombres, F., Brandt, M. M., Urbero, B., & Ayme, S. (2012). Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*. https://doi.org/10.1002/humu.22078

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, *11*, 95–130. https://doi.org/10.1613/jair.514

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*. https://doi.org/10.1038/gim.2015.30

Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., … Roden, D. M. (2013). Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation*. https://doi.org/10.1161/CIRCULATIONAHA.112.000604

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2008.09.017

Robinson, P. N., Köhler, S., Oellrich, A., Genetics, S. M., Wang, K., Mungall, C. J., … Smedley, D. (2014). Improved exome prioritization of disease genes

through cross-species phenotype comparison. *Genome Research*, *24*(2), 340–348. https://doi.org/10.1101/gr.160325.113

Robinson, P. N., & Mundlos, S. (2010). The Human Phenotype Ontology. *Clinical Genetics*, *77*(6), 525–534. https://doi.org/10.1111/j.1399-0004.2010.01436.x

Rosse, C., & Jr., L. V. M. (2007). The Foundational Model of Anatomy Ontology. *Esophagus*. https://doi.org/10.1007/978-1-84628-885-2_4

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3051

Sawyer, S. L., Hartley, T., Dyment, D. A., Beaulieu, C. L., Schwartzentruber, J., Smith, A., … Boycott, K. M. (2016). Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: Time to address gaps in care. *Clinical Genetics*. https://doi.org/10.1111/cge.12654

Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*. https://doi.org/10.1038/nmeth.2890

Schwarze, K., Buchanan, J., Taylor, J. C., & Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *GENETICS in MEDICINE*. https://doi.org/10.1038/gim.2017.247

Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P., & Musen, M. A. (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*, *10 Suppl 9*, S14. https://doi.org/10.1186/1471-2105-10-S9-S14

Shalita, A. R. (2004). Acne: Clinical presentations. *Clinics in Dermatology*. https://doi.org/10.1016/j.clindermatol.2004.03.012

Sifrim, A., Popovic, D., Tranchevent, L. C., Ardeshirdavani, A., Sakai, R., Konings, P., … Moreau, Y. (2013). EXtasy: Variant prioritization by genomic data fusion. *Nature Methods*. https://doi.org/10.1038/nmeth.2656

Simpson, M. A., Irving, M. D., Asilmaz, E., Gray, M. J., Dafou, D., Elmslie, F. V, … Trembath, R. C. (2011). Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. *Nature Genetics*, *43*(4), 303–305. https://doi.org/10.1038/ng.779

Singleton, M. V., Guthery, S. L., Voelkerding, K. V., Chen, K., Kennedy, B., Margraf, R. L., … Yandell, M. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *American Journal of Human Genetics*, *94*(4), 599–610. https://doi.org/10.1016/j.ajhg.2014.03.010

Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics*. https://doi.org/10.1038/ng1706

Smedley, D., Jacobsen, J. O. B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., … Robinson, P. N. (2015). Next-generation diagnostics and disease-gene

discovery with the Exomiser. *Nature Protocols*. https://doi.org/10.1038/nprot.2015.124

Smedley, D., Köhler, S., Czeschik, J. C., Amberger, J., Bocchini, C., Hamosh, A., … Robinson, P. N. (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics (Oxford, England)*, 1–8. https://doi.org/10.1093/bioinformatics/btu508

Smedley, D., Oellrich, A., Köhler, S., Ruef, B., Westerfield, M., Robinson, P., … Mungall, C. (2013). PhenoDigm: Analyzing curated annotations to associate animal models with human diseases. *Database*. https://doi.org/10.1093/database/bat025

Smedley, D., & Robinson, P. N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Medicine*. https://doi.org/10.1186/s13073-015-0199-2

Smedley, D., Schubach, M., Jacobsen, J. O. O. B., Köhler, S., Zemojtel, T., Spielmann, M., … Robinson, P. N. N. (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2016.07.005

Smith, C. L., & Eppig, J. T. (2012). The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mammalian Genome*. https://doi.org/10.1007/s00335-012-9421-3

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management CIKM 07*, *308*(1–2), 623. https://doi.org/10.1145/1321440.1321528

Sobreira, N., Schiettecatte, F., Valle, D., & Hamosh, A. (2015). GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. *Human Mutation*. https://doi.org/10.1002/humu.22844

Stenson, P. D., Mort, M., Ball, E. V, Howells, K., Phillips, A. D., Thomas, N. S., & Cooper, D. N. (2009). The Human Gene Mutation Database: 2008 update. *Genome Medicine*, *1*(1), 13. https://doi.org/10.1186/gm13

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J. B., Rimmer, A., … McVean, G. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*. https://doi.org/10.1038/ng.3304

The Deciphering Developmental Disorders Study. (2014). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, *10*(Chr X), 223–228. https://doi.org/10.1038/nature14135

The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1108

The Michigan Genomics Initiative. (2016). No Title. Retrieved August 6, 2018, from https://www.michigangenomics.org/

Topaz, M., Shafran-Topaz, L., & Bowles, K. H. (2013). ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. *Perspectives in Health Information Management / AHIMA, American Health Information Management Association*.

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., … Caulfield, M. J. (2018). The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *BMJ (Online)*. https://doi.org/10.1136/bmj.k1687

U.S National Library of Medicine. (2018). SNOMED CT. Retrieved June 14, 2018, from https://www.nlm.nih.gov/healthit/snomedct/index.html

UK Biobank. (2018). No Title. Retrieved August 6, 2018, from http://www.ukbiobank.ac.uk/

Van Der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*. https://doi.org/10.1007/s10479-011-0841-3

van Driel, M. a, Bruggeman, J., Vriend, G., Brunner, H. G., & Leunissen, J. a M. (2006). A text-mining analysis of the human phenome. *European Journal of Human Genetics : EJHG*, *14*(5), 535–542. https://doi.org/10.1038/sj.ejhg.5201585

van Zelst-Stams, W. A., Scheffer, H., & Veltman, J. A. (2014). Clinical exome sequencing in daily practice: 1,000 patients and beyond. *Genome Medicine*. https://doi.org/10.1186/gm521

Varadi, D. P., & Saqueton, A. C. (1970). Perifollicular Elastolysis. *British Journal of Dermatology*. https://doi.org/10.1111/j.1365-2133.1970.tb12875.x

Verma, A., Lucas, A., Verma, S. S., Zhang, Y., Josyula, N., Khan, A., … Pendergrass, S. A. (2018). PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2018.02.017

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2017.06.005

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq603

Weale, M. E. (2010). Quality control for genome-wide association studies. *Methods in Molecular Biology (Clifton, N.J.)*. https://doi.org/10.1007/978-1-60327-367-1_19

Wei, B., Pang, Y., Zhu, H., Qu, L., Xiao, T., Wei, H. C., … He, C. D. (2010). The epidemiology of adolescent acne in North East China. *Journal of the*

European Academy of Dermatology and Venereology. https://doi.org/10.1111/j.1468-3083.2010.03590.x

Wellcome, T., Case, T., & Consortium, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. https://doi.org/10.1038/nature05911

Wetterstrand, K. (2016). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Retrieved June 11, 2018, from www.genome.gov/sequencingcosts

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq340

Williams, H. C., Dellavalle, R. P., & Garner, S. (2012). Acne vulgaris. In *The Lancet*. https://doi.org/10.1016/S0140-6736(11)60321-8

Winand, R., Hens, K., Dondorp, W., De Wert, G., Moreau, Y., Vermeesch, J. R., … Aerts, J. (2014). In vitro screening of embryos by whole-genome sequencing: Now, in the future or never? *Human Reproduction*. https://doi.org/10.1093/humrep/deu005

World Health Organisation. (2010). *International statistical classification of diseases and related health problems*.

Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., Van Kogelenberg, M., … Firth, H. V. (2015). Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. *The Lancet*, *385*(9975), 1305–1314. https://doi.org/10.1016/S0140-6736(14)61705-0

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. a, … Eng, C. M. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *The New England Journal of Medicine*, *369*(16), 1502–1511. https://doi.org/10.1056/NEJMoa1306555

Yang, Y., Muzny, D. M., Xia, F., Niu, Z., Person, R., Ding, Y., … Eng, C. M. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*. https://doi.org/10.1001/jama.2014.14601

Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., … Robinson, P. N. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science Translational Medicine*. https://doi.org/10.1126/scitranslmed.3009262

Zondervan, K. T., & Cardon, L. R. (2007). Designing candidate gene and genome-wide case-control association studies. *Nature Protocols*. https://doi.org/10.1038/nprot.2007.366

Appendix 1

1.1 HPO terms for the patient with the MED12 nonframeshift deletion diagnosis (4.3.6) that wasn't captured by the exome filtering strategy employed here.

HPO code	**HPO term**
HP:0001274	Agenesis of corpus callosum
HP:0001631	Atria septal defect
HP:0001488	Bilateral ptosis
HP:0004325	Decreased body weight
HP:0002015	Dysphagia
HP:0001332	Dystonia
HP:0001508	Failure to thrive
HP:0001290	Generalized hypotonia
HP:0001263	Global developmental delay
HP:0001528	Hemihypertrophy
HP:0010864	Intellectual disability, severe
HP:0000252	Microcephaly
HP:0001561	Polyhydramnios
HP:0000407	Sensorineural hearing impairment
HP:0001518	Small for gestational age
HP:0001629	Ventricular septal defect

1.2 HPO terms for the patient with the CREBBP nonsynonymous SNV diagnosis (4.3.6) that wasn't captured by the exome filtering strategy employed here.

HPO code	**HPO term**
HP:0001999	Abnormal facial shape
HP:0000593	Abnormality of the anterior chamber
HP:0000481	Abnormality of the cornea
HP:0000359	Abnormality of the inner ear
HP:0002088	Abnormality of the lung
HP:0000587	Abnormality of the optic nerve
HP:0000356	Abnormality of the outer ear
HP:0000479	Abnormality of the retina
HP:0000069	Abnormality of the ureter
HP:0000795	Abnormality of the urethra
HP:0000925	Abnormality of the vertebral column
HP:0002251	Aganglionic megacolon
HP:0000062	Ambiguous genitalia
HP:0011675	Arrhythmia
HP:0001251	Ataxia
HP:0001631	Atria septal defect
HP:0007018	Attention deficit hyperactivity disorder
HP:0000717	Autism

HP:0000708	Behavioral abnormality
HP:0003561	Birth length less than 3rd percentile
HP:0100490	Camptodactyly of finger
HP:0001836	Camptodactyly of toe
HP:0005306	Capillary hemangiomas
HP:0001638	Cardiomyopathy
HP:0000518	Cataract
HP:0001396	Cholestasis
HP:0002072	Chorea
HP:0000175	Cleft palate
HP:0000204	Cleft upper lip
HP:0001680	Coarctation of aorta
HP:0000589	Coloboma
HP:0001674	Complete atrioventricular canal defect
HP:0000405	Conductive hearing impairment
HP:0000776	Congenital diaphragmatic hernia
HP:0001363	Craniosynostosis
HP:0000028	Cryptorchidism
HP:0004325	Decreased body weight
HP:0000750	Delayed speech and language development
HP:0000819	Diabetes mellitus
HP:0001332	Dystonia
HP:0002910	Elevated hepatic transaminases
HP:0002032	Esophageal atresia
HP:0001738	Exocrine pancreatic insufficiency
HP:0006101	Finger syndactyly
HP:0001371	Flexion contracture
HP:0001543	Gastroschisis
HP:0001290	Generalized hypotonia
HP:0001263	Global developmental delay
HP:0000085	Horseshoe kidney
HP:0000953	Hyperpigmentation of the skin
HP:0001010	Hypopigmentation of the skin
HP:0000047	Hypospadias
HP:0000601	Hypotelorism
HP:0002659	Increased susceptibility to fractures
HP:0010864	Intellectual disability, severe
HP:0009816	Lower limb undergrowth
HP:0000252	Microcephaly
HP:0000568	Microphthalmia
HP:0002011	Morphological abnormality of the central nervous system
HP:0000639	Nystagmus
HP:0001849	Oligodactyly (feet)
HP:0001180	Oligodactyly (hands)
HP:0001539	Omphalocele
HP:0001561	Polyhydramnios
HP:0100259	Postaxial polydactyly
HP:0004467	Preauricular pit
HP:0000384	Preauricular skin tag
HP:0100258	Preaxial polydactyly
HP:0000107	Renal cyst
HP:0002650	Scoliosis
HP:0001250	Seizures

HP:0000407	Sensorineural hearing impairment
HP:0002652	Skeletal dysplasia
HP:0001518	Small for gestational age
HP:0001257	Spasticity
HP:0010301	Spinal dysraphism
HP:0000486	Strabismus
HP:0001762	Talipes equinovarus
HP:0001636	Tetralogy of Fallot
HP:0001770	Toe syndactyly
HP:0002575	Tracheoesophageal fistula
HP:0009824	Upper limb undergrowth
HP:0011276	Vascular skin abnormality
HP:0001629	Ventricular septal defect
HP:0000505	Visual impairment

Appendix 2

Questions removed from the acne phenotype questionnaire.

Data “Type” column coded as follows:

- B: Binary
- Q: Quantitative
- F: Factor

“Reason” column coded as follows:

- M: Monomorphic
- I: Irrelevant
- T: Free-text (some binary columns are coded with the “free-text” reason for omission because they pertain to text entered in another column.
- P: Phenotypic information but irrelevant to acne
- F: Poorly filled

Name	*Type*	*%filled*	*Unique*	*Reason*
patientDetails: Forename	F	100	27	I
patientDetails: Surname	F	100	26	I
patientDetails: DateOfBirth	Q	100	5139	I
patientDetails: Ethnicity_Specify	F	3.4	54	M
familyHistory: Ps_FamilyHistory	B	0.3	4	P
familyHistory: Psoriasis	F	0.2	9	P
familyHistory: Acne	F	11	116	T
familyHistory: Ecz_FamilyHistory	B	0.3	3	P
familyHistory: Eczema	F	0.3	12	P
familyHistory: HS_FamilyHistory	B	8.6	4	P
familyHistory: Hidradenitis Suppurativa	F	0.5	16	P
diagnosis: Disease	F	100	1	I
diagnosis: Year_Diagnosis	Q	100	6	I
diagnosis: Age_Diagnosis	Q	0.1	6	I
diagnosis: Year_Onset	Q	99.9	60	I
acneDiagnosis: Disease	F	100	1	M
acneDiagnosis: Acne_Vulgaris	B	100	1	M

Name	Type	%filled	Unique	Reason
acneDiagnosis: Acne_EvidenceVirilisation	B	97.4	3	M
acneDiagnosis: Acne_Rosacea	B	98	3	M
acneDiagnosis: Acne_HormonalAbnormal	B	98	3	M
acneDiagnosis: Acne_Drug_Induced	B	98.1	3	M
acneDiagnosis: Acne_HH	B	98.1	3	M
acneDiagnosis: Acne_Occupation	B	97.8	3	M
acneDiagnosis: Acne_Occupation_Specify	B	0	1	M
acneDiagnosis: Acne_Other	B	45.2	4	I
acneDiagnosis: Acne_Other_Specify	F	6.9	32	T
acneDiagnosis: Acne_Isotretinoin	B	95.6	4	P
acneDiagnosis: Acne_LeedsScore	B	74.9	4	I
acneDiagnosis: Acne_YearMenarche	Q	24.8	56	I
acneDiagnosis: Acne_BloodTransfussion	B	100	1	M
clinical: Date_Clinical	Q	71.2	1708	I
clinical: Weight	Q	0.6	44	F
clinical: Height	Q	0.6	29	F
clinical: BMI	Q	0.6	46	F
clinical: Lesions_Site1	F	6.3	24	T
clinical: Lesions_Site2	F	0.7	16	T
clinical: Lesions_Site3	F	0	3	T
clinical: Comedones_Site1	B	1.7	3	T
clinical: Comedones_Site2	B	0.2	3	T
clinical: Comedones_Site3	B	0	2	T
clinical: Papules_Site1	B	3	3	T
clinical: Papules_Site2	B	0.4	3	T
clinical: Papules_Site3	B	0	2	T
clinical: Pustules_Site1	B	2.5	3	T
clinical: Pustules_Site2	B	0.3	3	T
clinical: Pustules_Site3	B	0	1	T
clinical: Cysts_Site1	B	2.2	3	T
clinical: Cysts_Site2	B	0.3	3	T
clinical: Cysts_Site3	B	0	2	T
clinical: LeedsScore_Site1	Q	2.3	13	T
clinical: LeedsScore_Site2	Q	0.3	6	T
clinical: LeedsScore_Site3	Q	0	1	T
clinical: Unknown_Chest	B	12.7	4	P
clinical: Unknown_Back	B	13.1	4	P
clinical: Unknown_Face	B	14.7	4	P
clinical: Unknown_Site1	B	1.7	3	P
clinical: Unknown_Site2	B	0.2	3	P
clinical: Unknown_Site3	B	0	1	P
clinical: Scarring_Site1	F	5.8	22	T
clinical: Scarring_Site2	F	0.6	15	T
clinical: Scarring_Site3	F	0.2	9	T
clinical: Scarring_General	B	0	1	M

Name	Type	%filled	Unique	Reason
clinical: Scarring_Site4	B	3.5	3	T
clinical: Scarring_Site5	B	0.2	3	T
clinical: Scarring_Site6	B	0.1	3	T
clinical: Hypertrophic_Site4	B	0.7	3	T
clinical: Hypertrophic_Site5	B	0.1	3	T
clinical: Hypertrophic_Site6	B	0.1	3	T
clinical: Keloid_Site4	B	0.5	3	T
clinical: Keloid_Site5	B	0.1	3	T
clinical: Keloid_Site6	B	0.1	2	T
clinical: Atrophic_Site4	B	0.8	3	T
clinical: Atrophic_Site5	B	0.2	3	T
clinical: Atrophic_Site6	B	0.1	3	T
clinical: IcePick_Site4	B	1.5	3	T
clinical: IcePick_Site5	B	0.3	3	T
clinical: IcePick_Site6	B	0.1	3	T
clinical: Perifollicular_Site4	B	0.5	3	T
clinical: Perifollicular_Site5	B	0.1	3	T
clinical: Perifollicular_Site6	B	0.1	2	T
clinical: PostInflam_Site4	B	1.3	3	T
clinical: PostInflam_Site5	B	0.3	3	T
clinical: PostInflam_Site6	B	0.1	3	T
clinical: _Summary	B	16.5	3	T
clinical: Scarring_Overall	F	44.5	76	T
clinical: Lesion_Overall	F	50.4	29	T
treatments: Disease	F	2.8	2	I
treatments: Therapy	F	2.8	4	I
treatments: TherapyType	F	2.8	4	I
treatments: Dose	Q	0	2	I
treatments: Frequency	Q	0	1	I
treatments: Date_Start	Q	2.4	82	I
treatments: Date_End	Q	0.3	22	I
treatments: OnGoing	B	1.6	4	I
treatments: Date_OnGoing	Q	1.1	69	I
*treatments: Responder**	B	2.1	5	I

DLQI: Dermatology Life Quality Index (Finlay & Khan, 1994).

*Although “Response to treatment” is interesting in the context of genetic studies of acne, there were so few responses in the entire treatments section of the questionnaire that all questions were removed.

Appendix 3

Questionnaire fields retained for analysis. *T* column is the data type (B: binary; Q: quantitative; F: Factor). U column is the number of unique responses. The final column displays the range and median for quantitative measures and the frequency of the most common response for other data types (excluding No/missing responses).

Name	*T*	*%F*	*U*	*Range (median) or Top%*
patientDetails: Sex	F	100	3	F: 56.6%
patientDetails: Ethnicity	F	100	3	White: 100.0%
familyHistory: Acne_FamilyHistory	B	18.8	3	Yes: 15.8%
diagnosis: Age_Onset	Q	79.4	38	1 - 41 (14)
acneDiagnosis: Acne_Infatile	B	20.6	3	Yes: 0.0%
acneDiagnosis: Assoc_PCOS	B	41.2	3	Yes: 1.0%
acneDiagnosis: Acne_Nodulocystic	B	56.4	3	Yes: 32.8%
acneDiagnosis: Acne_Fulminans	B	41.7	3	Yes: 0.3%
acneDiagnosis: Acne_Conglobata	B	41.8	3	Yes: 0.4%
acneDiagnosis: Acne_Sandpaper	B	42.3	3	Yes: 0.7%
acneDiagnosis: Acne_Submarine	B	42	3	Yes: 0.5%
acneDiagnosis: Acne_HairBeforeEleven	B	19.9	3	Yes: 3.3%
acneDiagnosis: Acne_BeforePuberty	B	68.8	3	Yes: 7.3%
acneDiagnosis: Hirsutism	B	1.2	3	Yes: 0.1%
acneDiagnosis: AndrogenicAlopecia	B	1.5	3	Yes: 0.0%
acneDiagnosis: Acne_HasBeenPregnant	B	10	3	Yes: 2.9%
acneDiagnosis: Acne_UnwantedFacialHair	B	9.9	3	Yes: 1.4%
clinical: DLQI	Q	27	32	0 - 30 (5)
clinical: Comedones_Chest	B	15.9	3	Yes: 7.9%
clinical: Comedones_Back	B	18.6	3	Yes: 11.2%
clinical: Comedones_Face	B	30.5	3	Yes: 25.2%
clinical: Papules_Chest	B	19.6	3	Yes: 12.1%
clinical: Papules_Back	B	24	3	Yes: 17.5%
clinical: Papules_Face	B	36.7	3	Yes: 32.3%
clinical: Pustules_Chest	B	16.7	3	Yes: 8.8%
clinical: Pustules_Back	B	20.7	3	Yes: 13.6%
clinical: Pustules_Face	B	31.6	3	Yes: 26.8%
clinical: Cysts_Chest	B	13.3	3	Yes: 4.6%
clinical: Cysts_Back	B	16.5	3	Yes: 8.6%
clinical: Cysts_Face	B	24.5	3	Yes: 18.6%
clinical: LeedsScore_Back	Q	21.8	14	0 - 12 (3)

Name	T	%F	U	Range (median) or Top%
clinical: LeedsScore_Chest	Q	19.5	13	0 - 11 (1)
clinical: LeedsScore_Face	Q	27.2	14	0 - 12 (5)
clinical: LeedsScore_Total	Q	28.3	34	0 - 43 (7)
clinical: Scarring_Chest	B	14.3	3	Yes: 3.5%
clinical: Scarring_Back	B	16	3	Yes: 5.6%
clinical: Scarring_Face	B	23	3	Yes: 14.2%
clinical: Hypertrophic_Chest	B	12.6	3	Yes: 1.1%
clinical: Hypertrophic_Back	B	13.1	3	Yes: 1.8%
clinical: Hypertrophic_Face	B	13.1	3	Yes: 1.8%
clinical: Keloid_Chest	B	12.2	3	Yes: 0.7%
clinical: Keloid_Back	B	12.3	3	Yes: 0.8%
clinical: Keloid_Face	B	12.1	3	Yes: 0.5%
clinical: Atrophic_Chest	B	13	3	Yes: 1.6%
clinical: Atrophic_Back	B	14.4	3	Yes: 3.2%
clinical: Atrophic_Face	B	20.2	3	Yes: 10.4%
clinical: IcePick_Chest	B	12.2	3	Yes: 0.7%
clinical: IcePick_Back	B	12.9	3	Yes: 1.4%
clinical: IcePick_Face	B	18	3	Yes: 8.3%
clinical: Perifollicular_Chest	B	11.9	3	Yes: 0.6%
clinical: Perifollicular_Back	B	12.5	3	Yes: 1.3%
clinical: Perifollicular_Face	B	12.5	3	Yes: 1.3%
clinical: PostInflam_Chest	B	13.6	3	Yes: 2.4%
clinical: PostInflam_Back	B	15	3	Yes: 4.4%
clinical: PostInflam_Face	B	19.1	3	Yes: 9.9%
clinical: Comedones_Summary	B	32.2	3	Yes: 27.2%
clinical: Papules_Summary	B	39.1	3	Yes: 35.2%
clinical: Pustules_Summary	B	34.6	3	Yes: 30.3%
clinical: Cysts_Summary	B	27.7	3	Yes: 22.3%
clinical: Scarring_Summary	B	27.2	3	Yes: 18.7%
clinical: Hypertrophic_Summary	B	14.3	3	Yes: 3.3%
clinical: Keloid_Summary	B	12.9	3	Yes: 1.5%
clinical: Atrophic_Summary	B	21.7	3	Yes: 12.0%
clinical: IcePick_Summary	B	18.9	3	Yes: 9.3%
clinical: Perifollicular_Summary	B	13.3	3	Yes: 2.4%
clinical: PostInflam_Summary	B	20.8	3	Yes: 12.0%

Appendix 4

Question response rates between individuals in population 1 and 2 identified in the missingness PCA (Figure 36, page 177) sorted by the difference in response rates, as well as the proportion of "No" responses for these questions.

	Pop1 fill	*Pop2 fill*	*rr diff*	*Pop1 no*	*Pop2 no*
clinical: Keloid_Face	0.005	0.9984	-0.9935	0.0833	0.989
clinical: IcePick_Chest	0.0068	1	-0.9932	0.0606	0.9859
clinical: Keloid_Chest	0.0068	0.9984	-0.9916	0.0303	0.9937
clinical: Perifollicular_Chest	0.0041	0.9953	-0.9912	0.0	0.9826
clinical: Keloid_Back	0.0083	0.9984	-0.9902	0.05	0.989
clinical: Hypertrophic_Chest	0.0105	1	-0.9895	0.0	0.9859
clinical: IcePick_Back	0.0138	1	-0.9862	0.0299	0.9796
clinical: Perifollicular_Face	0.0099	0.9953	-0.9854	0.0208	0.9621
clinical: Atrophic_Chest	0.0153	1	-0.9847	0.0135	0.978
clinical: Perifollicular_Back	0.0107	0.9953	-0.9845	0.0192	0.9685
clinical: Hypertrophic_Back	0.0161	1	-0.9839	0.0	0.9717
clinical: Keloid_Summary	0.0151	0.9984	-0.9833	0.0274	0.9796
clinical: Hypertrophic_Face	0.0169	1	-0.9831	0.0	0.9702
clinical: Perifollicular_Summary	0.0194	0.9953	-0.9759	0.0106	0.9401
clinical: PostInflam_Chest	0.0231	0.9953	-0.9721	0.0179	0.9653
clinical: Hypertrophic_Summary	0.0302	1	-0.9698	0.0	0.9466
clinical: Atrophic_Back	0.0308	1	-0.9692	0.0134	0.9529
clinical: PostInflam_Back	0.0393	0.9953	-0.956	0.0158	0.9164
clinical: Scarring_Chest	0.0324	0.9812	-0.9487	0.0064	0.9472
clinical: Scarring_Back	0.0513	0.9827	-0.9315	0.004	0.9058
clinical: IcePick_Face	0.0719	1	-0.9281	0.0029	0.8352
clinical: IcePick_Summary	0.0825	1	-0.9175	0.0025	0.8273
clinical: PostInflam_Face	0.0853	0.9953	-0.9099	0.0024	0.7918
clinical: Atrophic_Face	0.0973	1	-0.9027	0.0021	0.8477
clinical: PostInflam_Summary	0.1044	0.9953	-0.8909	0.002	0.7587
clinical: Atrophic_Summary	0.1137	1	-0.8863	0.0018	0.8289
clinical: Scarring_Face	0.1304	0.9859	-0.8555	0.0016	0.7643
clinical: Scarring_Summary	0.1775	0.9859	-0.8084	0.0012	0.7357
clinical: Comedones_Chest	0.0781	0.7692	-0.6911	0.1138	0.798
clinical: Cysts_Chest	0.0535	0.7363	-0.6827	0.2046	0.8977
clinical: Pustules_Chest	0.0884	0.7614	-0.6729	0.1028	0.8021
clinical: Comedones_Back	0.1087	0.7771	-0.6684	0.0722	0.7515
clinical: Papules_Chest	0.1201	0.7755	-0.6554	0.074	0.7449
clinical: Cysts_Back	0.0893	0.741	-0.6517	0.1042	0.8178
clinical: Pustules_Back	0.1312	0.7802	-0.649	0.0567	0.7022

	Pop1 fill	*Pop2 fill*	*rr diff*	*Pop1 no*	*Pop2 no*
clinical: Papules_Back	0.1684	0.7865	-0.6181	0.0368	0.6567
clinical: Cysts_Face	0.1734	0.7865	-0.6131	0.0381	0.5828
clinical: Comedones_Face	0.235	0.8367	-0.6018	0.0273	0.4822
clinical: Comedones_Summary	0.2529	0.8446	-0.5916	0.0212	0.4572
clinical: Cysts_Summary	0.2093	0.7928	-0.5834	0.0286	0.5267
clinical: Pustules_Face	0.2501	0.8148	-0.5647	0.0165	0.4701
clinical: Pustules_Summary	0.2819	0.832	-0.5501	0.0139	0.4075
clinical: Papules_Face	0.3048	0.8352	-0.5303	0.0129	0.4098
clinical: Papules_Summary	0.3306	0.8509	-0.5202	0.0106	0.3635
acneDiagnosis: Acne_Fulminans	0.3606	0.8462	-0.4855	0.9926	0.9963
acneDiagnosis: Acne_Conglobata	0.3614	0.8462	-0.4847	0.9874	0.9981
acneDiagnosis: Acne_Submarine	0.3643	0.8462	-0.4818	0.9836	1
acneDiagnosis: Acne_Sandpaper	0.3668	0.8477	-0.4809	0.9792	0.9944
acneDiagnosis: Acne_Nodulocystic	0.523	0.8728	-0.3498	0.3888	0.5558
clinical: DLQI	0.2385	0.5118	-0.2733	0.0737	0.0583
acneDiagnosis: Acne_BeforePuberty	0.659	0.9089	-0.2499	0.8968	0.8791
clinical: LeedsScore_Chest	0.1783	0.3203	-0.1419	0.3001	0.4118
diagnosis: Age_Onset	0.7783	0.9121	-0.1338	0.0	0.0
acneDiagnosis: Assoc_PCOS	0.397	0.5275	-0.1305	0.9776	0.9613
clinical: LeedsScore_Back	0.2038	0.3265	-0.1228	0.1886	0.3221
clinical: LeedsScore_Total	0.2689	0.3862	-0.1173	0.0077	0.0285
clinical: LeedsScore_Face	0.2585	0.3736	-0.1151	0.0208	0.0336
acneDiagnosis: Acne_HairBeforeEleven	0.1891	0.2732	-0.0841	0.8361	0.8391
familyHistory: Acne_FamilyHistory	0.1856	0.2041	-0.0185	0.1782	0.0385
acneDiagnosis: Hirsutism	0.0112	0.0188	-0.0077	0.9815	0.8333
acneDiagnosis: AndrogenicAlopecia	0.0143	0.0188	-0.0046	0.9855	1
patientDetails: Sex	0.9998	1	-0.0002	0.0	0.0
patientDetails: Ethnicity	1	1	0	0.0	0.0
acneDiagnosis: Acne_UnwantedFacialHair	0.1025	0.0754	0.0271	0.8629	0.8542
acneDiagnosis: Acne_HasBeenPregnant	0.1029	0.0754	0.0276	0.7129	0.6875
acneDiagnosis: Acne_Infatile	0.2277	0.0408	0.1869	0.9991	1

Appendix 5

Power calculation for each binary trait in the questionnaire, using numbers of controls for the original control dataset (C), "No" responses (N) and all non-"Yes" responses (N+X). Analyses were only performed if they had 80% power to detect a risk allele with 50% frequency and a genotype relative risk (GRR) of 2.0 (highlighted in bold; n=36). Effective cases and controls delineate the number of cases and controls after taking into account instances where DS1 or DS2 analyses were removed due to insufficient sample numbers.

Question	*Controls*	*Total cases*	*Total controls*	*DS1 cases*	*DS1 controls*	*DS2 cases*	*DS2 controls*	*Effective cases*	*Effective controls*	*GRR (80% power, 50% DAF)*
familyHistory: Acne_FamilyHistory	**C**	**863**	**21120**	**13**	**4976**	**850**	**16144**	**863**	**21120**	**1.4**
	N	863	165	13	2	850	163	850	163	2.25
	N+X	**863**	**4613**	**13**	**1734**	**850**	**2879**	**863**	**4613**	**1.4**
acneDiagnosis: Acne_Infatile	C	1	21120	0	4976	1	16144	0	0	>3
	N	1	1127	0	7	1	1120	0	0	>3
	N+X	1	5475	0	1747	1	3728	0	0	>3
acneDiagnosis: Assoc_PCOS	C	56	21120	15	4976	41	16144	56	21120	>3
	N	56	2201	15	208	41	1993	56	2201	>3
	N+X	56	5420	15	1732	41	3688	56	5420	>3
acneDiagnosis: Acne_Nodulocystic	**C**	**1794**	**21120**	**447**	**4976**	**1347**	**16144**	**1794**	**21120**	**1.25**
	N	**1794**	**1293**	**447**	**12**	**1347**	**1281**	**1794**	**1293**	**1.4**
	N+X	**1794**	**3682**	**447**	**1300**	**1347**	**2382**	**1794**	**3682**	**1.3**
acneDiagnosis: Acne_Fulminans	C	15	21120	3	4976	12	16144	12	16144	>3

Question	*Controls*	*Total cases*	*Total controls*	*DS1 cases*	*DS1 controls*	*DS2 cases*	*DS2 controls*	*Effective cases*	*Effective controls*	*GRR (80% power, 50% DAF)*
	N	15	2269	3	45	12	2224	12	2224	>3
	N+X	15	5461	3	1744	12	3717	12	3717	>3
	C	23	21120	5	4976	18	16144	18	16144	>3
acneDiagnosis: Acne_Conglobata	N	23	2265	5	45	18	2220	18	2220	>3
	N+X	23	5453	5	1742	18	3711	18	3711	>3
	C	40	21120	22	4976	18	16144	40	21120	>3
acneDiagnosis: Acne_Sandpaper	N	40	2275	22	45	18	2230	40	2275	>3
	N+X	40	5436	22	1725	18	3711	40	5436	>3
	C	29	21120	22	4976	7	16144	22	4976	>3
acneDiagnosis: Acne_Submarine	N	29	2273	22	45	7	2228	22	45	>3
	N+X	29	5447	22	1725	7	3722	22	1725	>3
	C	**178**	**21120**	**11**	**4976**	**167**	**16144**	**178**	**21120**	**1.95**
acneDiagnosis: Acne_HairBeforeEleven	N	178	911	11	37	167	874	178	911	2.1
	N+X	**178**	**5298**	**11**	**1736**	**167**	**3562**	**178**	**5298**	**1.95**
	C	**399**	**21120**	**41**	**4976**	**358**	**16144**	**399**	**21120**	**1.6**
acneDiagnosis: Acne_BeforePuberty	**N**	**399**	**3369**	**41**	**494**	**358**	**2875**	**399**	**3369**	**1.6**
	N+X	**399**	**5077**	**41**	**1706**	**358**	**3371**	**399**	**5077**	**1.6**
	C	3	21120	0	4976	3	16144	0	0	>3
acneDiagnosis: Hirsutism	N	3	63	0	21	3	42	0	0	>3
	N+X	3	5473	0	1747	3	3726	0	0	>3
	C	1	21120	0	4976	1	16144	0	0	>3
acneDiagnosis: AndrogenicAlopecia	N	1	80	0	19	1	61	0	0	>3
	N+X	1	5475	0	1747	1	3728	0	0	>3
	C	158	21120	0	4976	158	16144	158	16144	2.05

Question	*Controls*	*Total cases*	*Total controls*	*DS1 cases*	*DS1 controls*	*DS2 cases*	*DS2 controls*	*Effective cases*	*Effective controls*	*GRR (80% power, 50% DAF)*
acneDiagnosis: Acne_HasBeenPregnant	N	158	388	0	0	158	388	158	388	2.35
	N+X	158	5318	0	1747	158	3571	158	3571	2.05
acneDiagnosis: Acne_UnwantedFacialHair	C	75	21120	0	4976	75	16144	75	16144	2.7
	N	75	469	0	0	75	469	75	469	2.95
	N+X	75	5401	0	1747	75	3654	75	3654	2.75
clinical: Comedones_Summary	**C**	**1490**	**21120**	**151**	**4976**	**1339**	**16144**	**1490**	**21120**	**1.3**
	N	**1490**	**272**	**151**	**1**	**1339**	**271**	**1339**	**271**	**1.85**
	N+X	**1490**	**3986**	**151**	**1596**	**1339**	**2390**	**1490**	**3986**	**1.35**
clinical: Papules_Summary	**C**	**1928**	**21120**	**198**	**4976**	**1730**	**16144**	**1928**	**21120**	**1.25**
	N	**1928**	**214**	**198**	**1**	**1730**	**213**	**1730**	**213**	**1.95**
	N+X	**1928**	**3548**	**198**	**1549**	**1730**	**1999**	**1928**	**3548**	**1.3**
clinical: Pustules_Summary	**C**	**1659**	**21120**	**133**	**4976**	**1526**	**16144**	**1659**	**21120**	**1.3**
	N	**1659**	**235**	**133**	**1**	**1526**	**234**	**1526**	**234**	**1.9**
	N+X	**1659**	**3817**	**133**	**1614**	**1526**	**2203**	**1659**	**3817**	**1.3**
clinical: Cysts_Summary	**C**	**1223**	**21120**	**81**	**4976**	**1142**	**16144**	**1223**	**21120**	**1.3**
	N	**1223**	**295**	**81**	**0**	**1142**	**295**	**1142**	**295**	**1.85**
	N+X	**1223**	**4253**	**81**	**1666**	**1142**	**2587**	**1223**	**4253**	**1.35**
clinical: Scarring_Summary	**C**	**1024**	**21120**	**79**	**4976**	**945**	**16144**	**1024**	**21120**	**1.35**
	N	**1024**	**463**	**79**	**10**	**945**	**453**	**1024**	**463**	**1.7**
	N+X	**1024**	**4452**	**79**	**1668**	**945**	**2784**	**1024**	**4452**	**1.4**
clinical: Hypertrophic_Summary	**C**	**180**	**21120**	**11**	**4976**	**169**	**16144**	**180**	**21120**	**1.95**
	N	180	603	11	10	169	593	180	603	2.15
	N+X	**180**	**5296**	**11**	**1736**	**169**	**3560**	**180**	**5296**	**1.95**
clinical: Keloid_Summary	C	84	21120	3	4976	81	16144	81	16144	2.6

Question	*Controls*	*Total cases*	*Total controls*	*DS1 cases*	*DS1 controls*	*DS2 cases*	*DS2 controls*	*Effective cases*	*Effective controls*	*GRR (80% power, 50% DAF)*
	N	84	625	3	11	81	614	81	614	2.8
	N+X	84	5392	3	1744	81	3648	81	3648	2.65
	C	**658**	**21120**	**66**	**4976**	**592**	**16144**	**658**	**21120**	**1.45**
clinical: Atrophic_Summary	**N**	**658**	**529**	**66**	**10**	**592**	**519**	**658**	**529**	**1.7**
	N+X	**658**	**4818**	**66**	**1681**	**592**	**3137**	**658**	**4818**	**1.45**
	C	**508**	**21120**	**25**	**4976**	**483**	**16144**	**508**	**21120**	**1.5**
clinical: IcePick_Summary	**N**	**508**	**528**	**25**	**10**	**483**	**518**	**508**	**528**	**1.75**
	N+X	**508**	**4968**	**25**	**1722**	**483**	**3246**	**508**	**4968**	**1.55**
	C	131	21120	3	4976	128	16144	128	16144	2.2
clinical: Perifollicular_Summary	N	131	597	3	11	128	586	128	586	2.4
	N+X	131	5345	3	1744	128	3601	128	3601	2.2
	C	**657**	**21120**	**35**	**4976**	**622**	**16144**	**657**	**21120**	**1.45**
clinical: PostInflam_Summary	**N**	**657**	**482**	**35**	**11**	**622**	**471**	**657**	**482**	**1.75**
	N+X	**657**	**4819**	**35**	**1712**	**622**	**3107**	**657**	**4819**	**1.45**